Martin Perez, Javier

Providing Carrier Grade voice Services with Session Initiation Protocol

School of Electrical Engineering

Thesis submitted for examination for the degree of Master of Science in Technology.

Espoo 2.4.2014

Thesis supervisor:

Prof. Raimo Kantola

Thesis instructor:

M.Sc. Marko Luoma

Aalto University
School of Electrical
Engineering

AALTO UNIVERSITY
SCHOOL OF ELECTRICAL ENGINEERING

ABSTRACT OF THE
MASTER'S THESIS

Author: Martin Perez, Javier

Title: Providing Carrier Grade voice Services with Session Initiation Protocol

Date: 2.4.2014 Language: English Number of pages:9+67

School of Electrical Engineering
Department of Communications and Networking
Professorship: Communication Engineering Code: S-72

Supervisor: Prof. Raimo Kantola

Instructor: M.Sc. Marko Luoma

SIP is defined as a protocol that enables end-to-end voice calls as well as for establishing multiparty, multimedia communications in IP-based networks. Presently, SIP is the most widely deployed intra-carrier VoIP protocol but it is also extensively utilized within many carrier networks for transporting voice/video calls over short and long distances. For all of these reasons, SIP can lay a claim to being the global standard for software based voice communication over IP.

Furthermore, an important driving force for IP telephony is cost savings for consumers and corporations with large data networks. The high cost of long-distance and international voice calls presents both a challenge and an opportunity and must be taken into account. A significant portion of this cost originates from regulatory taxes imposed on long distance voice calls within the legacy networks. Such surcharges do not apply to long-distance circuit networks carrying data traffic; thus, for a given bandwidth, making a data call is much less expensive than a voice call. The objective of this research is to acknowledge SIP based communications as way to provide a better, reliable, cost effective, resource efficient and service flexible method for communications. The results will show certain vulnerabilities or weaknesses of the method, but also point solutions.

This thesis explains the VoIP/SIP based telephony network with call routing and admission control for real time traffic flows, also considering the priority usage perspective. To accomplish the main objectives, of proving the advantages of VoIP over traditional voice communications, we will analyze concepts such as Assured Services SIP, Multi-Level Precedence, admission control and bandwidth broker network elements. Moreover, we will touch Signaling System 7 with Session Border Controller as well as a small comparison to H.323 protocol.

Keywords: Voice over Internet Protocol, Quality of Service, Signalling, Session Initiation Protocol, Routing, Availability, Session Border Controller

Contents

Abstract ii

Contents iii

Abbreviations v

1 Introduction 1
- 1.1 Thesis structure . . . 1
- 1.2 Research problem and hypothesis . . . 1
- 1.3 Research questions . . . 2
- 1.4 Methodology . . . 2
- 1.5 Justification . . . 2
- 1.6 Background . . . 5

2 Protocols 7
- 2.1 VoIP Core Network . . . 7
  - 2.1.1 SIP . . . 7
  - 2.1.2 H.323 . . . 8
  - 2.1.3 Protocol Interworking (SIP-H.323) . . . 9
  - 2.1.4 VPN . . . 10
  - 2.1.5 NAT traversal . . . 10
- 2.2 PSTN network . . . 12
  - 2.2.1 SS7/ISDN . . . 13
  - 2.2.2 Interconnection (ISUP-SIP) . . . 15
  - 2.2.3 COPS (Common Open Policy Service) . . . 16
  - 2.2.4 ENUM (E.164 NUmber Mapping) . . . 16
- 2.3 X.25 Packet-Switched Network . . . 17
- 2.4 IMS/SIP overlay network . . . 18
  - 2.4.1 Planes framework . . . 18
  - 2.4.2 Network elements . . . 19
  - 2.4.3 PSTN to IMS interconnection . . . 20
- 2.5 Mobile network . . . 21
  - 2.5.1 SGSN, GGSN and HLR elements . . . 22
  - 2.5.2 Delay in 3G networks (due to signaling) . . . 23
  - 2.5.3 Handover cases . . . 23
- 2.6 Session border controllers . . . 24
  - 2.6.1 NAT comparison . . . 25
- 2.7 Binary SIP . . . 27

3 Service Level Management 29
- 3.1 SLA . . . 30
- 3.2 QoS . . . 31
  - 3.2.1 Call Admission Control (CAC) . . . 35
  - 3.2.2 Bandwidth Broker . . . 39

3.2.3 Multi-Level Precedence and Preemption . . . 40
3.3 Routing . . . 42
3.3.1 Dynamic routing . . . 43

4 Security **48**
4.1 Protocols . . . 49
4.1.1 SIPS . . . 49
4.1.2 SRTP . . . 49
4.1.3 ZRTP . . . 49
4.2 SIP Security mechanisms . . . 50
4.2.1 IPsec/TLS . . . 51
4.2.2 Security agreement mechanism . . . 52
4.3 Security Architecture . . . 52
4.3.1 Trusted Domain . . . 53
4.3.2 Trusted-but-Vulnerable Domain . . . 53
4.3.3 Untrusted Domain . . . 54
4.3.4 Signaling Security Architecture . . . 54
4.3.5 Media Security Architecture . . . 55
4.3.6 Application Security Architecture . . . 55

5 Availability and services **56**
5.1 Availability . . . 57
5.2 Services . . . 60
5.2.1 Instant Messaging and Presence . . . 60

6 Conclusions **62**

7 Future perspectives **64**

References **65**

Abbreviations

AAA	Authentication, Authorization and Accounting
AC	Application Controller
AES	Advanced Encryption Standard
AS	Application Server
ASCII	American Standard Code for Information Interchange
AGCF	Access Gateway Control Function
AGW	Access Gateway
ALG	Application Level Gateway
API	Application Programming Interface
B2BUA	Back-to-Back User Agent
BB	Bandwidth Broker
BCF	Base station Control Function
BGCF	Breakout Gateway Control Function
BGP	Border Gateway Protocol
BSS	Base station subsystem
CA	Certificate Authority
CAC	Call Admission Control
CGS	Carrier Grade Service
CS	Circuit Switched
CSCF	Call Session Control Function
CP	Control Plane
DCE	Data Circuit-Terminating Equipment
DES	Data Encryption Standard
DTE	Data Terminal Equipment
E2E	End to End
EGP	Exterior Gateway Protocol
ERC	Emergency Response Center
EU	European Union
FICORA	Finnish Communications Regulatory Authority
FTP	File Transfer Protocol
GGSN	Gateway GPRS Support Node
GSM	Global System for Mobile Communications
GPRS	General Packet Radio Service
HLR	Home Location Register
HSS	Home Subscriber Server
HTTP	Hypertext Transfer Protocol
I-BCF	Interconnect Border Control Function
ICE	Interactive Connectivity Establishment
IETF	Internet Engineering Task Force
IGP	Interior Gateway Protocol
IM	Instant Messaging

IMS	Internet Multimedia Subsystem
ISDN	Integrated Services Digital Network
IS-IS	Internmediate System to Intermediate System
ISP	Internet Service Provider
ISUP	ISDN User Part
ITU	International Telecommunication Union
IWF	Interworking Function
IXC	Inter-eXchange Carrier
LAN	Local Area Network
LSA	Link State Advertisement
M3UA	MTP Level 3 User Adaptation
MCU	Multipoint Control Unit
MIME	Multipurpose Internet Mail Extensions
MGCP	Media Gateway Control Protocol
MGCF	Media Gateway Controller Function
MGW	Media Gateway
MIH	Media Independent Handover
MLPP	Multi-Level Precedence and Pre-emption
MS	Mobile Station
MSC	Mobile Switching Center
MTP	Message Transfer Part
MTU	Maximum Transmission Unit
NAT	Network Address Translation
NRDB	Network Resource Database
OAM	Operation, Administration and Maintenance
OSPF	Open Shortest Path First
PATS	Public Available Telephone Service
PBAS	Precedence-Based Assured Service
PBX	Private Branch eXchange
PKI	Public Key Infrastructure
PS	Packet Switched
PSE	Power Sourcing Equipment
PSTN	Public Switched Telephony Network
QoS	Quality of Service
RAU	Routing Area Update
RNC	Radio Network Controller
RTP	Real-time Transport Protocol
SBC	Session Border Controller
SCCP	Signalling Connection Control Part
SCTP	Stream Control Transmission Protocol
SCP	Service Control Point
SDP	Session Description Protocol
SGW	Signalling Gateway

SGSN	Serving GPRS Support Node
SIP	Session Initiation Protocol
SIPS	SIP Secure
SLA	Service Level Agreement
SLM	Service Level Management
SMTP	Simple Mail Transfer Protocol
SP	Service Provider
SRTP	Secure RTP
SS7	Signalling System 7
SSP	Service Switching Point
STP	Signalling Transfer Point
STUN	Session Traversal Utilities for NAT
TCAP	Transaction Capabilities Application Part
TCP	Transmission Control Protocol
TLS	Transport Layer Security
TRIP	Telephony Routing over IP
TURN	Traversal Using Relay NAT
UA	User Agent
UDP	User Datagram Protocol
UMTS	Universal Mobile Telecommunications System
UTRAN	UMTS Terrestrial Radio Access Network
VoIP	Voice over Internet Protocol
VPN	Virtual Private Network
WAN	Wide Area Network

List of Figures

1 Big picture SIP VoIP key concepts . . . 1
2 SIP SP access scenario with SBC . . . 8
3 SIP SP access scenario with SBC . . . 9
4 NAT traversal scenario . . . 11
5 NAT access scenario with SBC . . . 12
6 PSTN architecture . . . 13
7 Architecture of SS7 . . . 14
8 SS7 functional levels . . . 15
9 ISUP-SIP interconnection scenario . . . 16
10 X.25/X.21 scenario . . . 18
11 IMS network topology . . . 19
12 Mobile telephony network structure . . . 22
13 GPRS CN interfaces . . . 23
14 Vertical handover . . . 24
15 SBC architecture . . . 25
16 Call parts with SBCs . . . 26
17 NAT vs. SBC . . . 27
18 Interface between customer and service provider processes . . . 30
19 DiffServ QoS building blocks . . . 34
20 CAC . . . 36
21 SIP CS, SBC and PS . . . 37
22 UMTS IP interworking . . . 38
23 Bandwidth Broker . . . 40
24 Security protocols . . . 49
25 Security domains . . . 53

List of Tables

1 Priority rating specification 58
2 Battery back-up time 59

1 Introduction

1.1 Thesis structure

The main purpose of the present master thesis is to analyze, explain and evaluate all the pertaining areas that involve the SIP communications and VoIP transmissions. Looking to establish IP communications as a reliable and successful way to complement or, in some cases, completely take the place of legacy communications.

There will be 7 chapters in total; covering all the different aspects of SIP based transmissions. In Figure 1, we will be discovering some of the key concepts that will be discussed in this thesis. With SIP being the central piece of this thesis, we will analyze all the branches that SIP protocol touches.

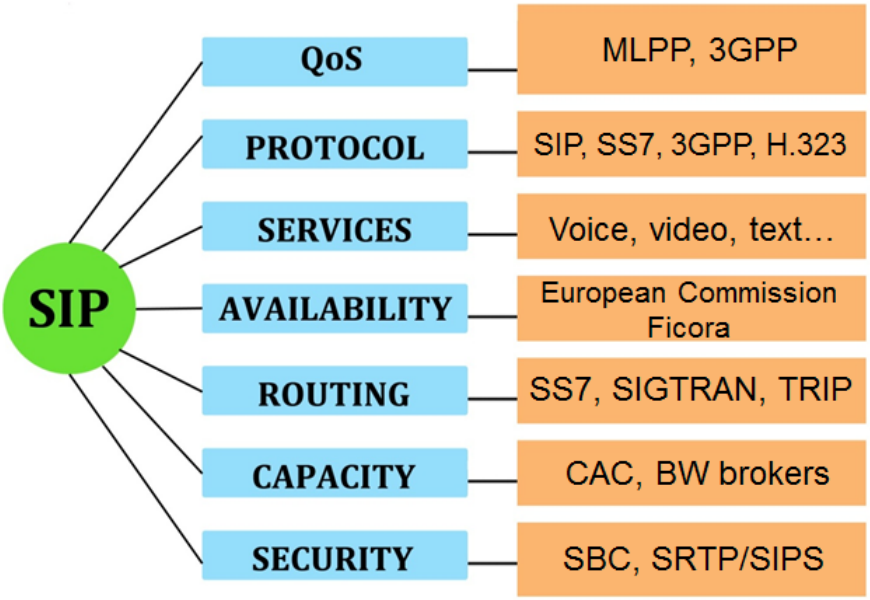


Figure 1: Big picture SIP VoIP key concepts

1.2 Research problem and hypothesis

Basing the software efforts on SIP, we will demonstrate that SIP is one of the best options for services over the IP network.

Legacy networks like PSTN will be studied and compared to more modern IP networks. Certainly, both kinds of networks rely on hardware components, but IP ones also utilize software heavily to optimize not only voice, but multimedia services.

We will focus on the SIP protocol as a base to establish a reliable communication service. As such, we can focus on the possibility to use SIP with VoIP communications and to provide the necessary services and security assurances for a good

alternative to conventional communications.

Along the way, and contained in the chapters, we will come to know many of the advantages that IP networks offer. We will describe what 'Carrier Grade' services are and the different problems to achieve it. 'Carrier Grade' will make waves only in so far as to the reliability of a service.

Most components of a network, either hardware or software related, have certain weaknesses or bottlenecks that must be addressed to comply with a functioning reliable service. 'Carrier Grade' mostly means that a process work under a 99.999% of availability. All the components have to work together in order to provide that high level of availability performance. Many of the discussed topics will not directly address the 'Carrier Grade' definition, but most solutions have to do with a better protection of continuous services and indirectly aiming for the 5 nines rule of 99.999% availability. [36]

1.3 Research questions

- How do packet-switched networks compare to circuit-switched ones?
- What are the security implications of IP Networks?
- What software solutions are needed for software based communications?
- What is the best protocol for VoIP communications?
- Are there different services that can be offered through IP networks?
- What is the reliability and availability of VoIP communications?
- What are the benefits and disadvantages of SIP based communications?

1.4 Methodology

The thesis will report a literature study and take a theoretical approach when describing the processes for IP communications. When talking about regulations, we will focus more on the European Commission's rules and the Finnish national regulator FICORA. Also, we will be discussing the future of SIP, and the different networks that may help the technology grow.

1.5 Justification

The major motivations is to demonstrate how SIP based communications can take a bigger role and become a more mainstream method, not only for voice services, but for media related data. As any other technology, SIP communication have certain priorities and some shortcomings, but with every passing year, technologies advance and weaknesses disappear. SIP communications can present a certain challenge as

more variables are involved that the hardware or power supply. SIP communications rely in a big and important software component that adds up to the previous communications technologies. Regardless, it will evolve and become a more reliable service aiming even for a 'Carrier Grade' status.

All the different components of a network, either software or hardware related, form a coordinated mechanism that can then be transformed into reliable services like VoIP. The term 'Carrier Grade' refers mostly to a qualitative adjective that defines if a network is reliable and what can any user and provider expect from it. It normally grades certain aspects like availability, backup reliability, quality of services provided among other features. [36]

In the past, electricity or power supply used to be one if not the main dilemma when dealing with voice communications. Hardware malfunction was also very important, as it will always had a tear and wear variability depending on the conditions it performs. As new technologies and networks have come to be born, many more variables get into the mix, sometimes making the 'Carrier Grade' rule of 99.999% difficult if not impossible to achieve.

This term has certainly a base in the hardware and software components and what kind of predictability you can give to any given customer. This term normally equals a 99.999% (five nines) of availability of a given system. For a service provider, this translates generally into a monetary investment in order to deliver high reliability, fast recovery from possible failures; not only on the hardware side of the infrastructure but also by having and improving the necessary protocols to manage and direct data traffic flow in the most optimal way.

First, we have to realize that many IP networks rely on the Internet to work. Internet as a mean to provide services has certain limitations as it does not comply with a 'Carrier Grade' policy, but a 'Best efforts' one, which does not offer any minimal level of service or availability. 'Carrier Grade' emphasises the requirement of availability, so that service providers can offer certain assurance in the service delivered. 'Carrier Grade', in plain words, may be described as the guarantee offered by communication service providers to consumers that the service delivered will fulfill the accorded parameters of availability performance, robustness and quick recover from failure. There is no such a thing as an infallible communication system, as there can be hardware, software, operator errors additionally to malicious and unintended errors form 3rd parties that may cause problems with the system functionality. Unlike legacy networks that mostly deal with the first 3 problems, IP networks have to deal with all of these aforementioned weaknesses. Delivery of this guarantees is plainly difficult for any service provider and thus, and even so, 'Carrier Grade' may not be even enough to provide an improved communication over IP system.

If we state that we are going to approach the topic mostly from a theoretical

point of view, we also assume to aim for local or closed networks that can deliver the promises of 'Carrier Grade' reliability. When describing non-local or open networks, it is not possible to aim for 'Carrier Grade' services. Furthermore, and this is not the main objective of this thesis, there are certain economical implications to consider, as most customer are not willing or prepared to pay for 'Carrier Grade' services. Service providers have to evaluate and balance their returns compared to the investment needed to provide better availability.

As such, some might argue that cost is an important variable to consider when deciding on the availability factor, and even if that's the case, it will be assumed that most of the solutions for many of the SIP VoIP communications will be aimed to better this goal even though some scholars agree that the five nines are sometimes unattainable. [36]

Some next generation IP networks will have or target 'Carrier Grade' operability, depending on the criteria utilized or the need of a certain customer, but most of the present IP networks still operate under the 'Best Effort' principle. Certainly, we will analyze the interoperability among many kinds of networks, and some of them, have to comply with 'Carrier Grade' quality, this represents a problem when most of the networks presently do not work under such high standards. Legacy networks, IP networks (mostly Ethernet ones) or mobile networks can work together to provide the most optimal communication experience according to certain minimal requirements that will be analyzed in this document.

The requirements in areas like software protocols, service level management, security, availability and quality of services, all comprise the most basic and comprehensive needs for a communication service. While describing, studying and analyzing every one of these areas, we can then start to conceive the overall idea of a 'Carrier Grade' communication system, in this specific case VoIP.

Certainly, we will try to describe the theoretical capacities of the VoIP technology and its many components, but at certain point we will have to also understand certain minimal specification required by the European Commission and the Finnish regulator FICORA. This will prove to be key within our understanding of the elements of a 'Carrier Grade' network, as mostly every aspect and detail will be based on this high demanding seal of quality.

Content wise, we will touch on all the related networks and its different structures and functions. Most of the world is still using the more conventional communications network or Public Switched Telephone Networks (PSTN). This refers to the circuit-switched network infrastructure that provides the normal telephone communication also known as the Plain Old Telephone Service. Thanks to the new IP network infrastructure provided by many service providers and the growth of the IP telephony, many companies can offer telephony services through this network. This is called IP Telephony or VoIP.

Currently, voice communication networks can be divided among the following categories:

- Analog Telephone Networks
- ISDN
- 2G and 3G mobile networks
- Lastly, IP telephone networks form ISP (Internet)

Every network mentioned above has its unique and independent infrastructure, equipment and support. One of the specific purposes of this research is to explain and describe these networks, their differences, how they interact and connect with each other. The new VoIP services like Skype, use mainly the IP network to keep their cost low, but also interact with all of the mentioned networks.

1.6 Background

Circuit Switching vs. Packet Switching

To compare the new VoIP services, we need to at least describe to some small extent the most common mediums of communication used today. Here, we can define the methods of communication between two big fields: the circuit switching and the packet switching.

During the circuit-switching mode, the course or path that the data will use is defined even before the transmission is started. Resource optimizing systems programed beforehand and utilizing certain algorithms, decide the best possible route. One of the main benefits of this kind of communication is that its exclusive and dedicated only for the purpose of this transmission and for its whole duration, which means that the resources used are only available again when the transmission ends. The units of data transmitted with circuit switching are not identified by any address, which means that it has a constant connection capacity 64 kb/s. Afterwards, a switching fabric state is established between and incoming circuit and an outgoing circuit.

The most important advantage a user gets by transmitting a signal over the circuit-switching mode is that it does not have to deal with any packets of data and do not have to solve the problems of the traffic that can occur in packet switching. The user also enjoys from an uninterrupted transfer of signal that wont need to deal with any bandwidth availability issues or resource delay. The connection is established end to end with a bidirectional connection. The most common disadvantage for this kind of communication is that in itself it is unsuitable for many kinds of

applications to use many kinds of applications as opposed to internet-related technology. In other words, it lacks the flexibility that many application need nowadays.

As examples of circuit switched networks we have:

- Public Switched Telephone Network (PSTN).
- ISDN B-channel.
- Circuit Switched Data (CSD) and High-Speed Circuit-Switched Data (HSCSD) service in cellular systems such as GSM.
- X.21 (Scandinavian DATEX circuit switched data network).

During a packet-switching transmission, the packets of information are sent towards the destination, usually through a certain defined route, depending on the agreements, policies and link weight; and the most efficient course at the time. They can be connection oriented or connectionless. For certain kinds of packets, certain end to end virtual connections must be established. Both communication technologies will be described in more detail in chapter 2.

2 Protocols

When speaking about the software related to any VoIP transmission, there are many kinds of protocols that help in the overall process. They can be used to provide a certain level of Quality of Service, to create a framework for multimedia or telephony gateways, to be able to receive certain kinds of texts or body headers, to initiate a signaling, to transport data in a certain way, to control and configure links, among many other things.

A protocol is series of rules and requirements that operate to exchange different kinds of information with different types of electronic devises.

This chapter is designed to introduce and define the most common and needed protocols for establishing a call with VoIP.

2.1 VoIP Core Network

2.1.1 SIP

This protocol is key among the many other ones, as it is the one that we are using to define the way we are going to work with VoIP communication during this project. Developed by the IETF for use in IP networks, the Session Initiation Protocol (SIP) is a signaling protocol with the main task of establishing, modifying and terminating transmissions among devises, users and switches. Regardless of its importance, it relies on many other transmission protocols like the UDP, TCP or SCTP. This protocol is ready to be used with both IPv4 and IPv6.

A characteristic of the SIP protocol is that it copies certain elements already in use with the HTTP and SMTP protocols. SIP works with UTF-8 text based messages, both for requesting and answering a message. Similar to HTTP, SIP borrows its client server design and utilizes URLs and URIs. Additionally, in the same way that SMTP encodes text and headers, SIP copies this and incorporates it to its layout.

SIP is a very developer friendly protocol as it is easily readable, due to its text message nature. The architecture is expandable and new features can be added easily according to the client and network's needs. [27]

The SIP entities are the following:

- User Agent (UA): It resides in a SIP end device.
- Proxy: It works on behalf of a user by responding to requests or forwarding them. A proxy server is not required to understand the full content of a message in order to redirect and does not change the order of header fields.

- Redirect server: It receives SIP requests and notifies the source of the original request on the location of the destination.
- Register and Location: It takes in SIP REGISTER requests and provides location services used by SIP redirect or proxy servers to be able to access information about possible locations of the callee.

Certain SIP gateways are used to interconnect to ISUP or H.323. When we are communicating between a packet-switched network (IP), and circuit-switched network (PSTN), gateways can terminate a signaling or media if necessary. When using a gateway with SIP and H.323, it does not necessarily end media, but the signaling does need to be processed.

SIP SP to Internet

The Session Border Controller (SBC) is placed at the border between the access network (IP core, outer network) and the operator's network (SIP SP, inner network). Figure 2 shows where the SBC would be located in a SIP scenario. The SBC is managed by the organization maintaining the operator network. [2]

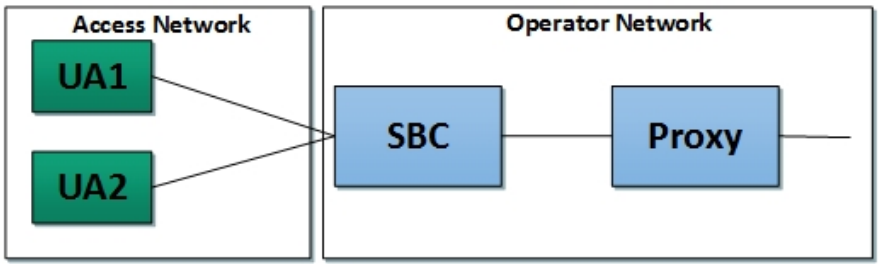


Figure 2: SIP SP access scenario with SBC

2.1.2 H.323

H.323 is not a protocol per se, but it encompasses a set of recommendations published by the ITU Telecommunication Standardization Sector (ITU-T). It does contain and specifies certain protocols and architectural structure of an IP multimedia system. It can be regarded as a guideline from the ITU-T.

A service provider manages this by dividing the network into administrative domains. When working with a H.323 network and functional entities, one central entity manages the whole operation. An administrative domain can be formed by any entity managed by the service provider.

H.323 entities can be described as:

- Terminals: it is an entity that terminates signaling and media at an end- user's location.
- MCU: It is in charge of administering multipoint conferences, which means, when two or more end-point devises are engaged in a conference. It is composed of two controllers, one for signaling, the MCU, which contains a Multipoint Controller (MC) and manages the call signaling and, Multipoint Processors (MPs), to handle media processing. H. 323 terminals and multi-point control units (MCU) terminate both signaling and media. They can be addressed as endpoints.
- Gatekeeper: This entity provides services to end users and routes messages to their final destinations. These include authentication, authorization and accounting (AAA) and address resolution.

H.323 signaling can also accomplish codec negotiations to obtain an appropriate codec between endpoints.

H.323 can communicate with other kinds of networks, like H.324, which is widely utilized by 3G networks, or H.320 (ISDN), via certain media gateways. This is an important aspect of the use of H.323 as mobile network use will only increase in the long term. [1]

2.1.3 Protocol Interworking (SIP-H.323)

This protocol is used, in our case, to interconnect SIP and H.323. This is achieved thanks to an InterWorking Function (IWF) that allows different systems with different signaling protocols to work together. It synchronizes thefunctions and information in different kinds of networks. The following figure 3 shows how this interconnection works between SIP and H.323 through an SBC with an InterWorking Function.

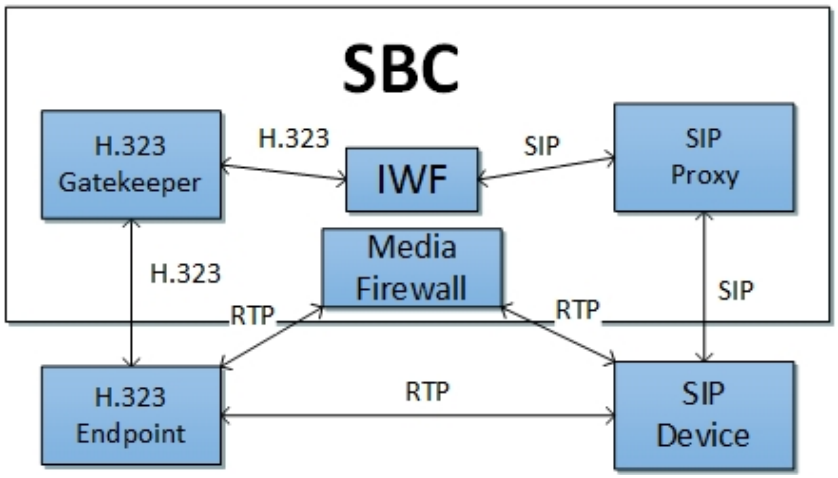


Figure 3: SIP SP access scenario with SBC

Protocol Interworking is also important, due to the fact that it may be necessary even if the two systems are working with the same signaling protocol. For example, between SIP services on the Internet and SIP based IMS services, there is an incompatibility within the same protocol and consequently, different SIP profiles must be used for each one of them. Processing to convert different signaling protocols is a very resource intensive task, as it requires much time and computing power to achieve it.

2.1.4 VPN

This refers to a Virtual Private Network (VPN) and it provides remote access to a specific central organizational network. This can be useful in order to have offices in different geographical regions for travelling users. In the case of VoIP, VPN delivers a secure method to provide voice services.

A VPN works with the normal VoIP services, but then encrypts the data packets through a VPN tunnel. The complete process includes converting the analog voice signal to a digital one, divide this into data packets, encrypt the message with IPsec, and the use of the VPN to secure the transmission. At the place where the user is accessing the data, the voice is then converted to an analog form to be listened by another user.

A VPN is also a useful way to avoid firewall problems when configuring remote SIP VoIP clients. The protocol overhead initiated by the encapsulation in data packets of VoIP protocol within IPSec increases the bandwidth requirements for VoIP calls, consequently making the VoIP over VPN protocols too big to be transferred by certain mobile data connections like GSM or UMTS. This means, that VoIP over VPN is not as usable in mobile networks, but nonetheless, it is sometimes used to create encrypted VoIP trunks between different sites of some corporations, running VoIP PBX interconnections over a VPN.

2.1.5 NAT traversal

Most destination terminals in a VoIP communication system, only have a private IP address and are then hidden behind a NAT. Unfortunately, SIP does not support full NAT traversal. To access private networks, NAT traversal mechanisms are needed, in order for the traffic to flow between private and public networks. To address this inconvenience, a NAT box must be used with the following elements: [11]

SBC

This feature can act as Back- to-Back User Agent (B2BUA) when talking about media, and at the same time, from the signaling point of view it can also be viewed as a SIP Proxy. Classic SBC has some disadvantages.

Inserting a SBC element, from the media point of view this element acts as Back-to-Back User Agent (B2BUA) while from the signaling point of view it can also act as a SIP Proxy. SBC needs no modification on existing network. However, classic SBC system faces serious problems as follows.

- SBC can become a bottle-neck: SIP proxies only redirect signaling while SBCs are in charge of media and signaling forwarding, This transforms into a traffic imbalance between the SIP proxies and the SBCs.
- Hard to extend: If traffic increases and a SBC overloads, the only way to solve the problem is to upgrade it with a more robust and high performance node.
- SBC opens a backdoor in the firewall of the enterprise: This leads to a potential problem to the security of the intranet. This is illustrated on the Figure 4, where we can look at a NAT traversal scenario using SBC network elements.
- Some endpoints are behind enterprise or residential NATs: When the user accesses a private network, the SBC is a NAT for all traffic. In Figure 5, a NAT scenario is graphically described and where a SBC should be located.

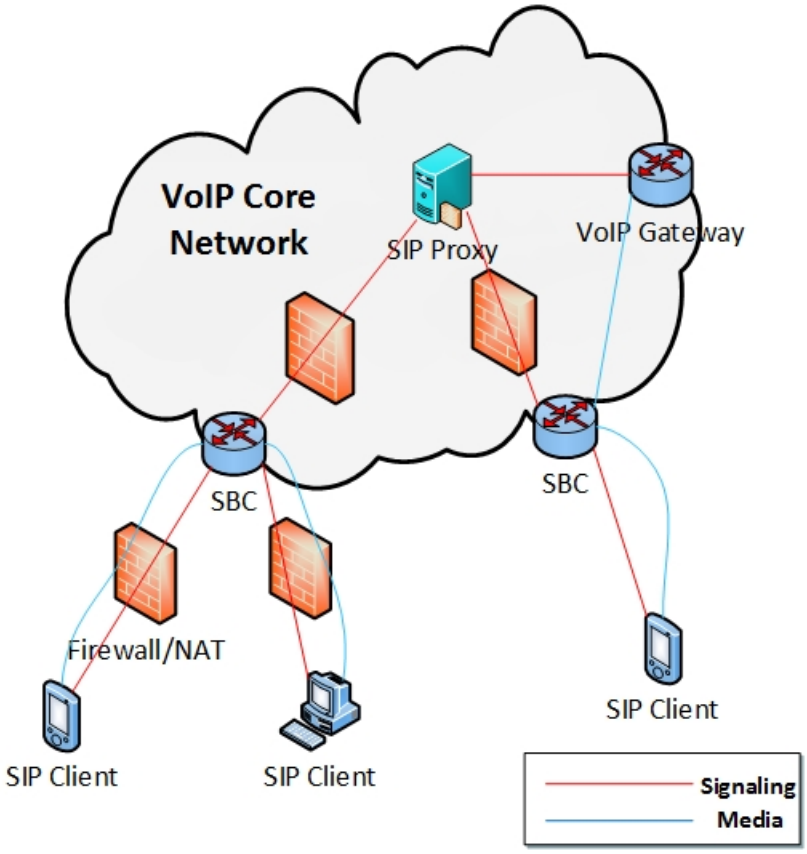


Figure 4: NAT traversal scenario

Some endpoints may be behind enterprise or residential NATs. In cases where the access network is a private network, the SBC is a NAT for all traffic. Figure 5

shows where the SBC would be located in a NAT scenario.

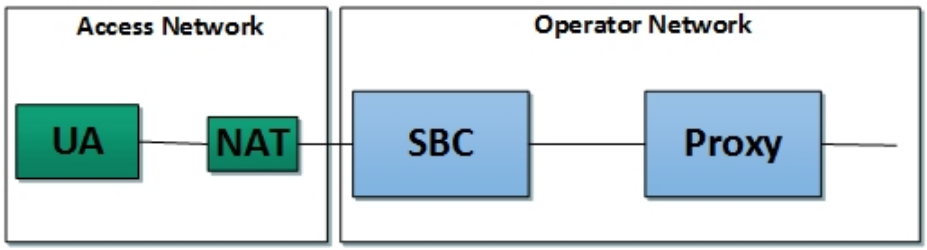


Figure 5: NAT access scenario with SBC

IETF protocols: STUN, TURN and ICE

STUN (Session Traversal Utilities for NAT) lets the different applications discover the public IP address and port mappings that they can use to communicate with their peers.

TURN (Traversal Using Relay NAT) as opposed to STUN, allocates a public IP/port on a global server and relays media among the parties involved in the communication.

ICE (Interactive Connectivity Establishment) helps to solve the important NAT traversal incompatibilities. It uses both STUN and TURN and chooses the best way to interconnect users and networks. As ICE includes the use of STUN and TURN, the complete solution can be addressed only as ICE.

Application Level Gateway (ALG)

This specific application, controls the flow of data for some other applications. Unfortunately, it is difficult to implement and is not 100% secure.

Generally, it use a method utilized for certain applications that use layer payload to communicate the dynamic TCP (Transmission Control Protocol) or the UDP (User Datagram Protocol) ports. Many IP protocols including the FTP (File Transfer Protocol) require the use of ALGs. These protocols are also called pinholes, and are only in use during the data transmission.

2.2 PSTN network

The public switched telephone network (PSTN), or traditional network, is composed by circuit-switched transmission with each stream using a 64 kb/s digital channel. This network can be either digital or analogue. Figure 8 shows the PSTN architecture with its main network elements. The PSTN nodes are described as:

- Signaling Transfer Point (STP): It performs message routing, transferring of signaling messages.
- Service Switching Point (SSP): It is the telephone exchange. When a telephone caller dials a number, this sends a query to a central database (SCP) so that the call can be correctly managed and redirected.
- Service Control Point (SCP): It receives requests from the SSP, and then uses service logic created in SCE (Service Creation Environment).
- Service Data Point (SDP): It stores all the user information necessary for other entities to make decisions. It serves as a central database.

Figure 6 shows the PSTN architecture with its main network elements.

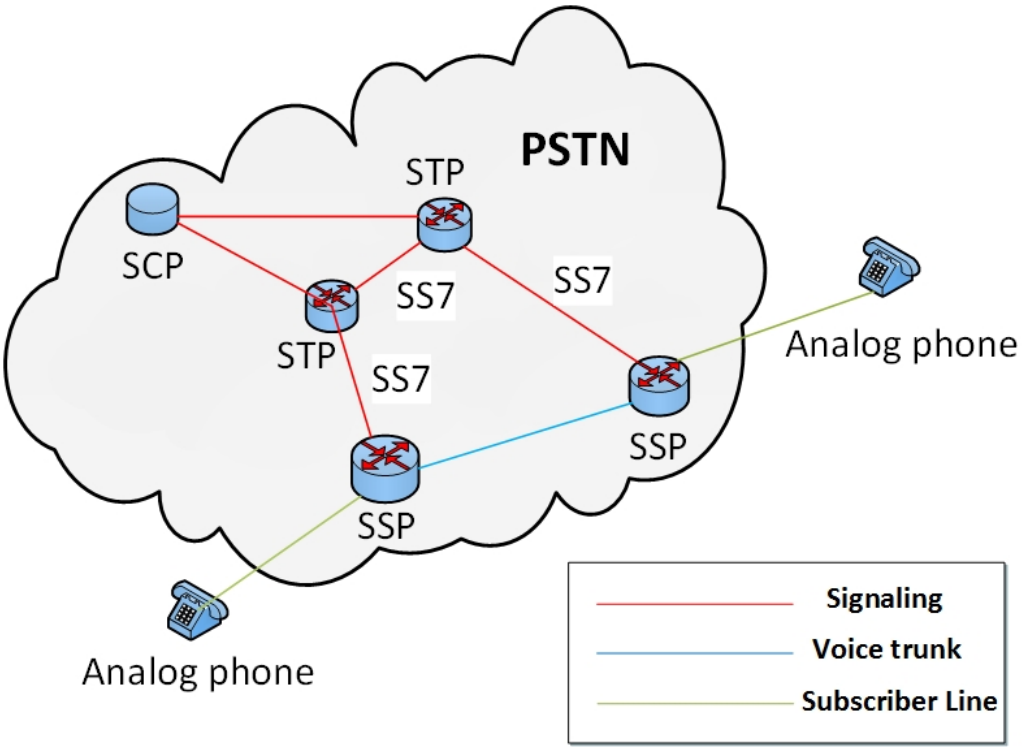


Figure 6: PSTN architecture

2.2.1 SS7/ISDN

SS7 (Signaling System no. 7) is the most commonly used signaling system. It allows for the signaling to be transmitted on a different logical channel than the call channel. The PSTN functions based on this signaling system which supports call control, routing, billing and information exchange functions.

Figure 7 illustrates the architecture of SS7 and the relationship among the various functional blocks of the SS7 and between its levels and the TCP/IP Reference

Model Layers.

As stated before, PSTN allows for digital and analogue connections, ISDN (Integrated Services Digital Network) allows these networks to support digital access. Figure 8 shows the functional levels of SS7. The main responsibility of the MTP (Message Transfer Part) is to provide a reliable transfer of signaling messages between the locations of communicating user functions. This refers to level 1-3. [10]

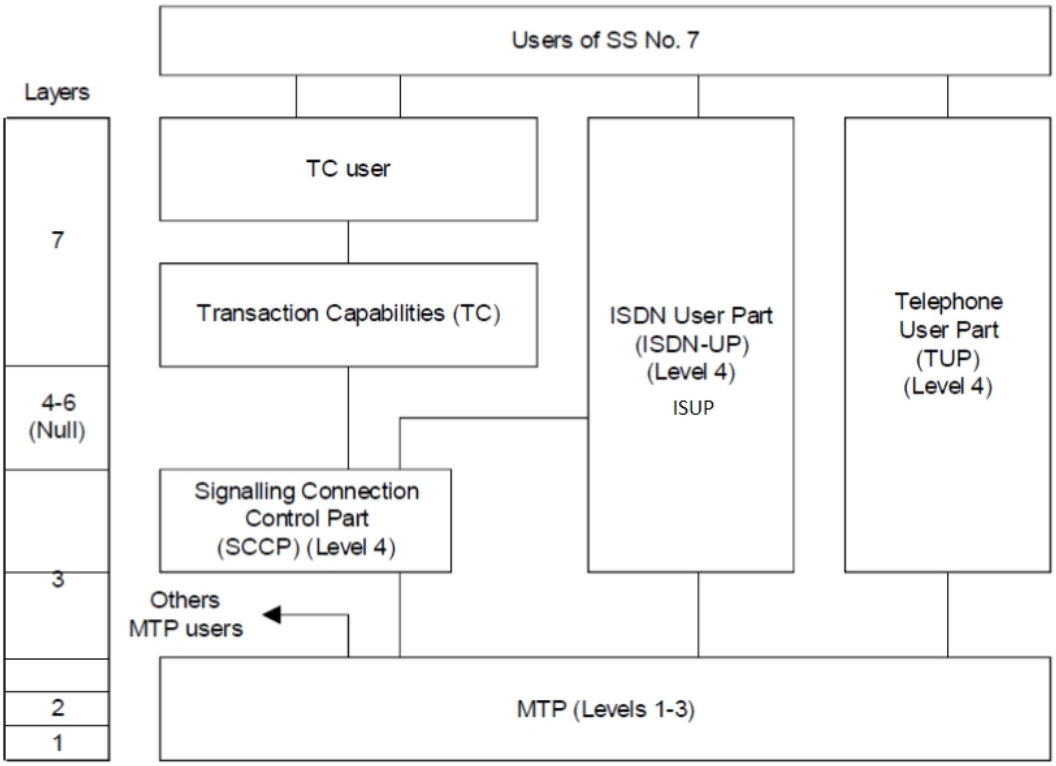


Figure 7: Architecture of SS7

Level 4 consists of the different User Parts:

- Signaling Connection Control Part (SCCP): It adds features to the MTP in order to supply connectionless and connection oriented services. It also supports circuit and non-circuit related signaling.
- Telephone User Part (TUP): It refers to the international telephone call control signaling functions for use over MTP.
- Data User Part (DUP): It settles the appropriate protocol to control circuits needed on data calls, data call facility registration and cancellation.
- ISDN User Part (ISUP): The ISUP gather the needed signaling features to deliver switched services and user facilities for voice and non-voice applications

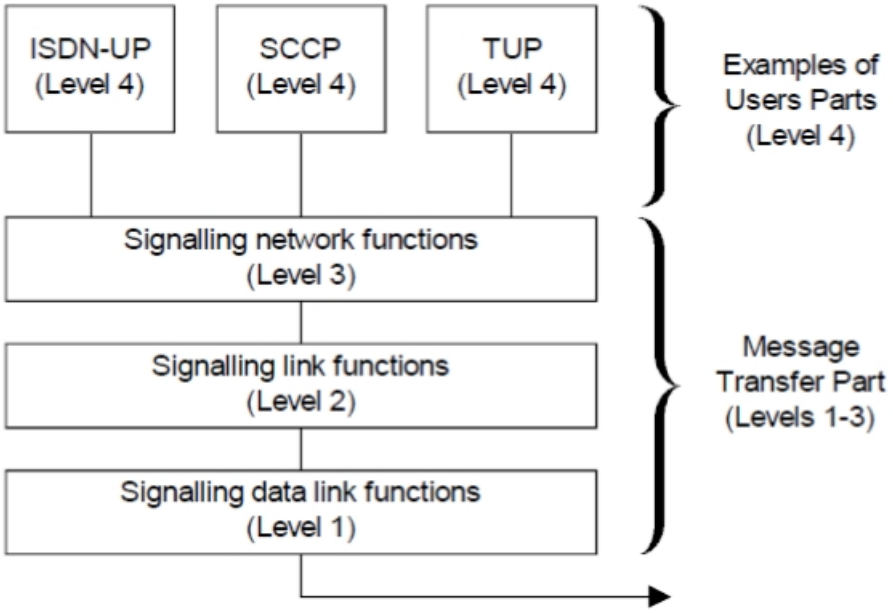


Figure 8: SS7 functional levels

in the ISDN. It is prepared to function with circuit-switched networks, meaning traditional telephone networks, but also with analogue/digital mixed ones. The ISUP has an interface to the SCCP to permit the ISUP to use the SCCP for end-to-end message delivery control.

- Transaction Capabilities (TC): It allows a non-circuit related communication between two different nodes.
- Applications: This are basically the contact between a system and the end user of the telephone or ISDN network, and the service provider to interact with supplementary services.

2.2.2 Interconnection (ISUP-SIP)

With Figure 9, we can appreciate a real interacting scenario between ISUP and SIP with some information related to the networks elements as well as the protocols interworking used for it.

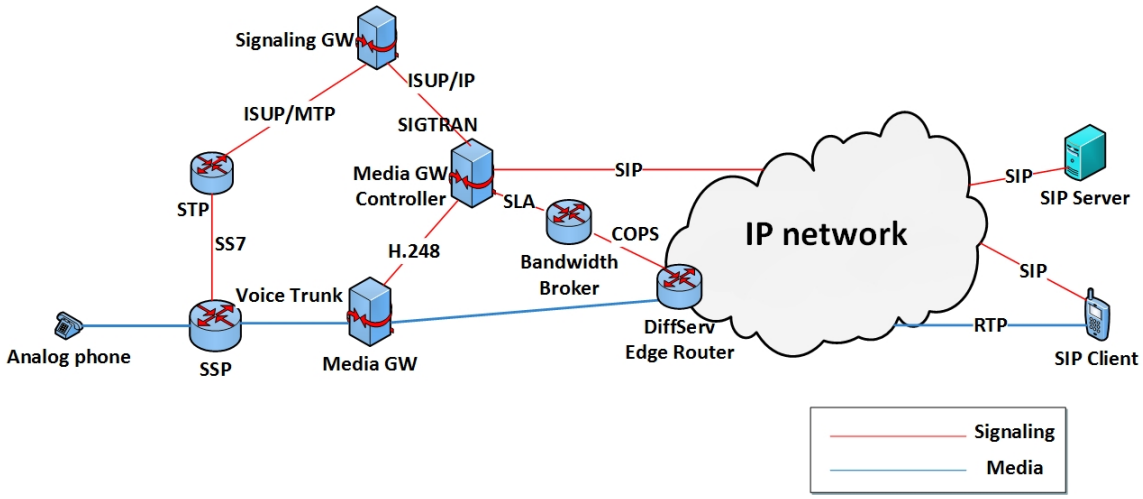
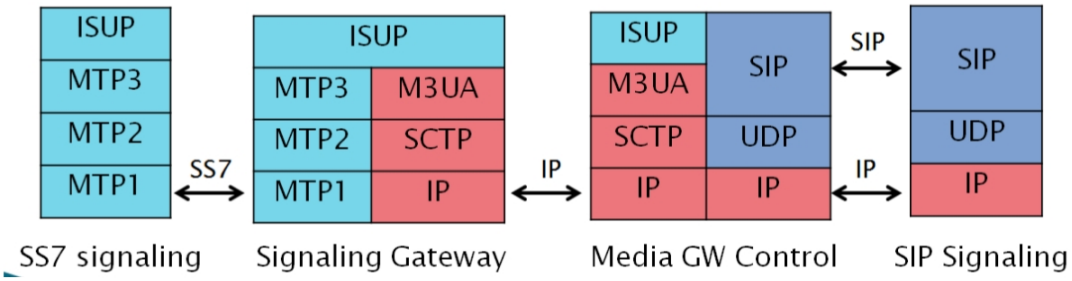


Figure 9: ISUP-SIP interconnection scenario

M3UA

M3UA is the acronym for MTP Level 3 (MTP3) User Adaptation Layer as defined by the IETF SIGTRAN (RFC 4666). M3UA allows the SS7 protocol user parts to be able to work over IP instead of traditional telephony equipment like ISDN and PSTN and it involves SCTP to transmit M3UA.



2.2.3 COPS (Common Open Policy Service)

COPS is a protocol that is mostly utilized for Quality of Service requirements. It basically exchanges policy information for a specified network between the PDP (policy decision point) and PEP (policy enforcement point). Everything always according to the priorities agreed beforehand. All these exchanges are made for the purpose of traffic allocation and prioritization.

2.2.4 ENUM (E.164 NUmber Mapping)

ENUM protocol was developed by the IETF (Internet Engineering Task Force) (RFC 3761) to solve the problem of how certain VoIP systems can find each other on the Internet while depending only on a fixed telephone number, but also vice versa, as

telephone can also access Internet services.

MGCP

MGCP (Media Gateway Control Protocol) as its name implies, controls media gateways on both IP networks and PSTN. It is a protocol that utilizes a PSTN over IP model that allows to signal and control calls with IP systems that interoperate with telephone lines.

H.248

Another Gateway Control Protocol used for media is Megaco. The IETF defined Megaco as something equal to H.248. As another type of media gateway protocol, it controls the media gateway between IP and PSTN networks. The only big difference between MGCP and H.248 is that it supports more commands, processes and more networks.

To control the communication between a Media Gateway Controller and the Gateway that converts circuit-switched voice data packets traffic, most systems make use of H.248 that is the standard to deliver VoIP when using both IP networks and the PSTN.

2.3 X.25 Packet-Switched Network

A packet-switched network refers always to a digital one. It collects all transmitted data, voice or media in our VoIP case, into defined data packets. A packet network, can be connection oriented like X.25 but also connectionless like IP networks. Every network that deals with data packets is shared with other networks and it transmits every packet independently, as opposed to PSTN that sends all at once without delving into too much efficiency. The transmission of the packets depends on the resources available.

As such, packet switching's main objective is to optimize the use of resources like link availability, decrease response times and better communication. Depending on other factors within the shared networks, the packet running through a certain route can be buffered or queued depending on the traffic load on the network.

To fulfill this task, we can use X.25 protocol as it is designed for communication on packet switched WAN (Wide Area Networks). The hardware is made of PSE (Packet Switching Exchange) nodes, but also comprises PSTN or ISDN connections. The X.25 specification defines the DTE (Interface Between a Subscriber) and an X.25 network (DCE). A packet switched network is illustrated in Figure 10.

X.25 worked as a packet switched service, but its architecture was thought to be similar to that of the circuit-switched old networks but relying on software to

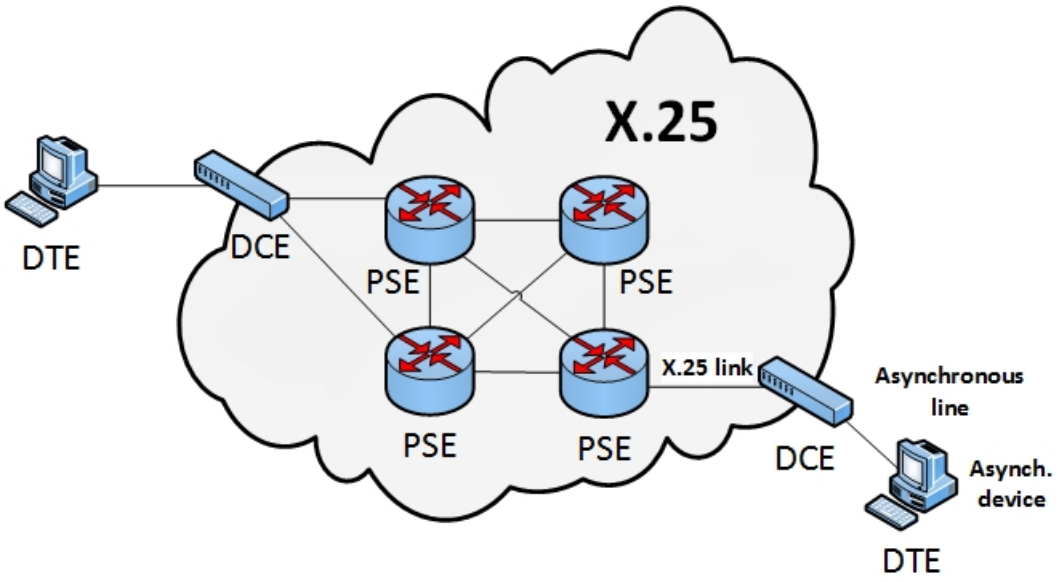


Figure 10: X.25/X.21 scenario

create virtual calls. A benefit of this virtual environment is that each end point can originate more than one call and DTE is interconnected providing point-to-point transmissions.

2.4 IMS/SIP overlay network

The IMS (IP Multimedia Subsystem) is an IP-based communication system in accordance with the 3GPP (3rd Generation Partnership Project). It permits the connection of fixed connections with cellular phone ones.

A special characteristic contained in IMS is that it can be configured independent from the access networks. This allows access to it from packet communications domains, from mobile 3G systems, wireless LAN and fixed networks. Its features also consist of authentication and session control. [37]

2.4.1 Planes framework

On the control plane, we use SIP to process the IP multimedia signaling. We use SDP to transport the media information. These are completely separated from the transportation layer which is packet switched based. To comprehend an IMS network topology better, Figure 11 shows the different planes in which a network performs.

IMS has the interoperability necessary to communicate with fixed, mobile and IP networks.

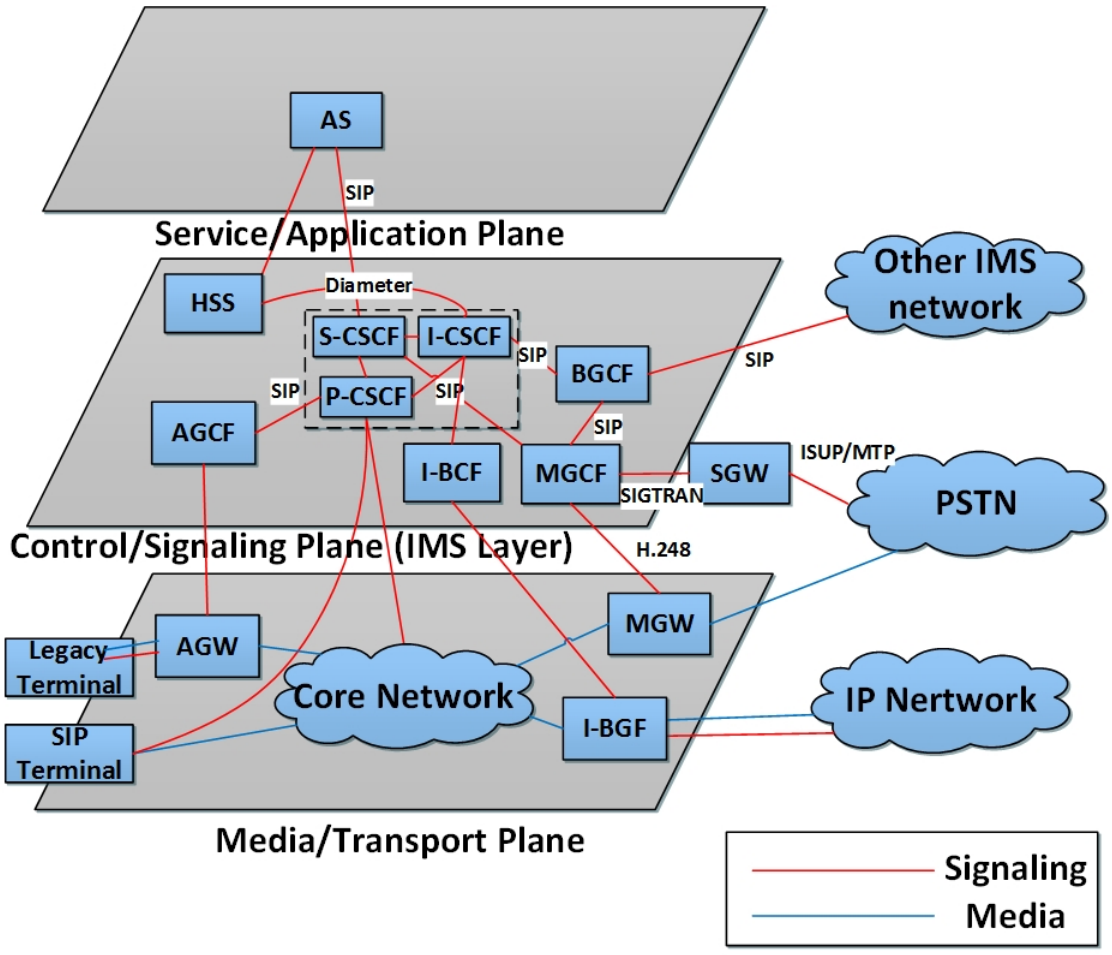


Figure 11: IMS network topology

2.4.2 Network elements

The IMS has many work planes as stated above, which are the Transport Plane, Control Plane and the Application Plane. To use SIP to signal among them we need the CSCF (Call Session Control Function).

P-CSCF (Proxy-CSCF) is an interface used to have a user-to-network interoperability and it's the contact provided for the IMS terminal. P-CSCF manages all the signaling that comes and exits from the IMS. Every IP address of a P-CSCF must be discovered first, in order for and IMS end-point to register. By collecting data about all the registered end points already in the network the P-CSCF secures the access to IMS network for access.

The S-CSCF (Serving-CSCF) acts as a central hub to the core network and requests all the services needed by using SIP signaling to the Application Servers, and deciding on a destination (routing).

The I-CSCF (Interrogating-CSCF) passes setup for the sessions from the initiating user 'A' S-CSCF to the destination user 'B' S- CSCF. It also is the entry point to any IMS network for transmission coming from outside the network.

The Home Subscriber Server (HSS), as it name implies, works as a deposit for all the subscriber data that the IMS network needs for services it provides, voice or multimedia. DIAMETER is a protocol used to communicate from the HSS to the IMS network by authenticating the subscriber. DIAMETER is now used instead of the old COPS protocol.

Security and policy control are also needed within the system, and for that purpose the I-BCF (Interconnection Border Control Function) is used. This helps by providing SBC on the SIP stream and then the I-BGF (Interconnect Border Gateway Function) applies a control to the RTP media stream.

MGCF, MGW, and SGW are used to connect traditional PSTN to the IMS Networks. The Breakout Gateway Control Function (BGCF) performs the routing of signaling requests to other networks. IMS Networks examine the identity of a caller party and determine if the call comes from within or outside the network.

During the processing of an originating call, the IMS network analyzes the called party ID to determine whether it corresponds to an on-net or off-net destination. An originating Application Server or an S-CSCF can perform this processing.

PSTN emulation refers to how a legacy network is connected to an IMS network. Fixed networks endpoints are TDM-based and the emulation subsystem work with 3 different elements to bring them together.

The Access Gateway (AGW) supports physical connectivity to TDM-based legacy terminals; the Access Gateway Control Function (AGCF) controls the AG and supports SIP signaling towards the IMS core.

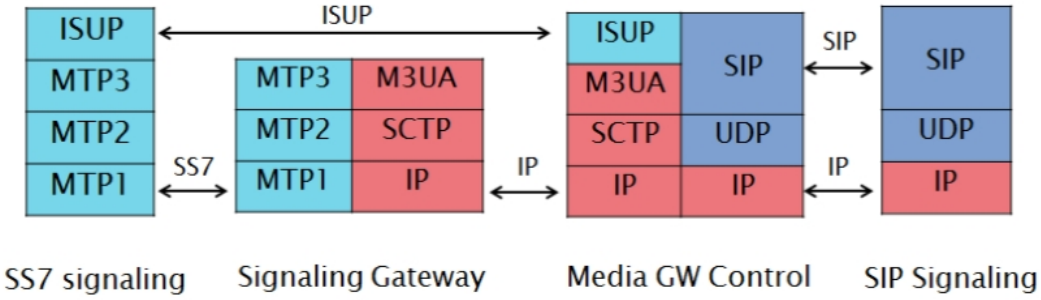
SIP Application Servers (SIP-AS) host and provide services, and interfaces with the S-CSCF. Different services might operate with different modes. The AS can operate in SIP proxy mode, SIP UA (User Agent) mode or SIP B2BUA mode. Depending on where the AS is physically located, the home network may require HSS for a SIP-AS.

2.4.3 PSTN to IMS interconnection

In circuit networks the ISUP protocol to manage the signaling for the call control.

SS7 is widely popular in the legacy networks, which includes Transaction Capabilities Application Part (TCAP), Signaling Connection Control Part (SCCP), and Message Transfer Part (MTP) protocols.

The following Figure illustrates the idea of interworking between IMS and legacy network. SCTP and M3UA protocols are used to support the transportation, and the SS7 signaling over IP and MGCF performs mapping between ISUP and SIP.



Voice bearers of IMS have to be connected with the voice bearers of circuit networks. The MGW is provided to support the bearer interworking and also has the E1/T1 interface directed to a PSTN and the IP interface for VoIP in IMS. It may support transcoding between a codec used by the UE in IMS and the codec used in the network of the a third party.

SGW also works by supplying the transport layer protocol conversion on which the MGCF works and handling the interworking in the signaling plane. The most critical feature or characteristic is performing SS7 over IP to help the routing of ISUP messages.

2.5 Mobile network

The UMTS (Universal Mobile Telecommunications System) is separated into the:

- Core Network (CN) infrastructure: It includes a CS (Circuit Switched) domain, a PS (Packet Switched) domain and an IP Multimedia Subsystem according to the 3GPP specifications
- Access Network (AN) infrastructure: The AN infrastructure includes the RNS (Radio Network Subsystem) for UMTS also known as UMTS Terrestrial Radio Access Network (UTRAN) and the GSM Base Station Subsystem (GSM BSS).

The UTRAN and the GSM networks are the ones in charge of radio communications; CN manages the switching and routing signaling and data transmission to the outer networks. GPRS CN is the network connecting the last two. This structure is present in the Figure 12 that shows a mobile telephony network.

The GSM Base Station Subsystem (GSM BSS) accommodates radio access, when a Mobile Switching Center (MSC) is doing the circuit switching. A Home Location Register (HLR) using a mobile specific protocol (MAP) that tracks down the call.

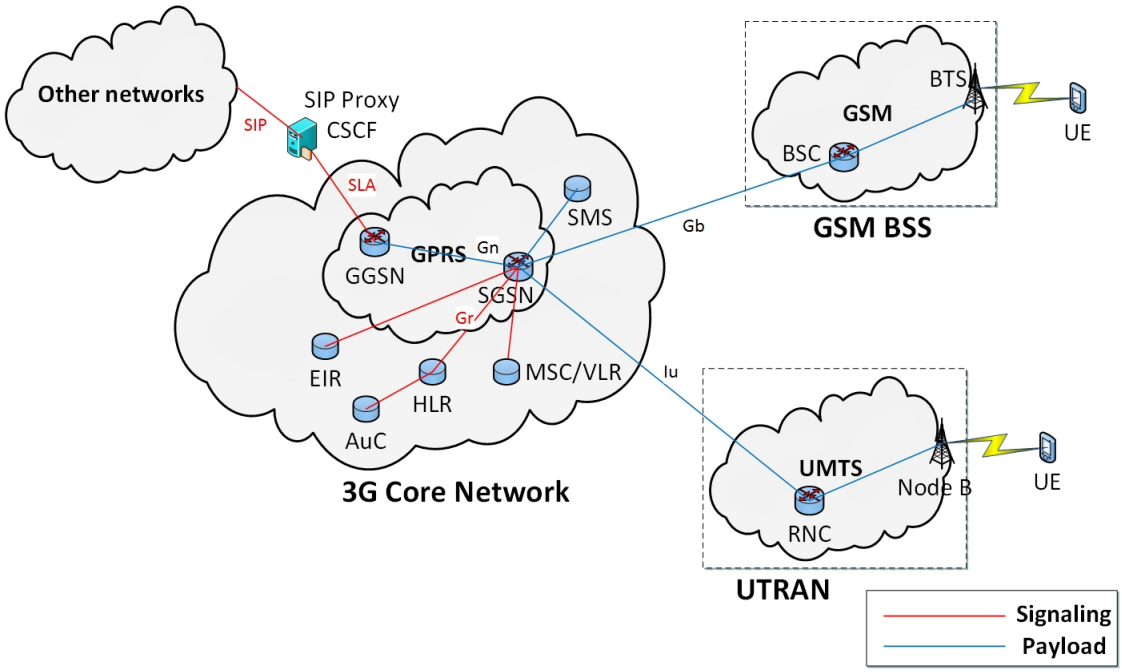


Figure 12: Mobile telephony network structure

The General Packet Radio Service (GPRS) recycles the GSM radio interface (GSM BSS) for packet-service support but uses new network components for routing, the SGSN and GGSN.

2.5.1 SGSN, GGSN and HLR elements

The core network for packet switched services includes tree main nodes:

- SGSN: It refers to a very important component in the UMTS core network. The SGSN redirects incoming and outgoing IP packets directed to/from a mobile station that is connected within the SGSN service area. This SGSN has also mobility management, session management, authentication and charging responsibilities.

- GGSN: This is regarded as the external packet data networks and the gateway between the mobile core networks. It has a similar duty as an Internet router. Data packets between SGSN and GGSN are sent with GPRS Tunneling Protocol (GTP). The most important objective of GGSN is to tunnel and de-tunnel the packets coming from or destined to the mobile station. GGSN also deals with access control, session management and charging responsibilities.

- HLR: It can be defined as a database containing subscription data for each of the subscribers. There may be more than one HLR nodes in any given network. The SGSN requests information from the HLR in order to receive subscription, quality of service and other related data.

2.5.2 Delay in 3G networks (due to signaling)

For better schematics of the reason why there is delay in 3G networks, we must define the main interfaces between the network elements, as Figure 13 illustrates:

- Iu: This connects the SGSN to RNCs, permitting the exchange of signaling and payload. The Iu Control plane interface is connected by means of SS7. The Iu User plane interface works using IP, and it is only needed in 3G systems for connecting the radio network to the core network.
- Gn: This allows the exchange of signaling and payloads by connecting the SGSN to other SGSNs and to GGSNs in the same Public Land Mobile Network (PLMN) using IP.
- Gr: This connects the SGSN to the HLR by means of SS7 allowing subscriber data management.

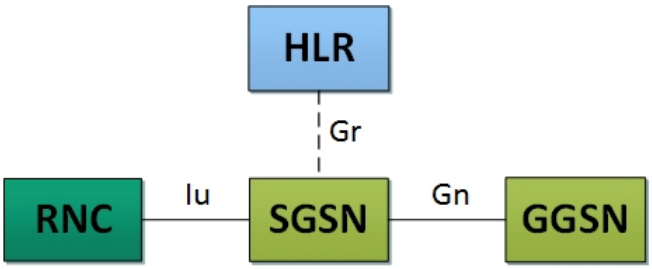


Figure 13: GPRS CN interfaces

Operations made by the 3G core network:

Inter-SGSN RAU (Routing Area Update) and another SGSN attach are two of the single operations causing the largest signaling load. Inter-SGSN routing area updates results in a significant part of the signaling traffic in the Gr interface with the SGSN and the HLR. [11] The share of signaling traffic compared to the total traffic can amount to almost 20 per cent within the core network. [8]

2.5.3 Handover cases

There are two types of handover:

- Vertical handover refers to the automatic change from one technology to another in order to continue the communication.
- Horizontal handover occurs between different wireless access points that use the same technology. It does not change like vertical handover to access the network.

Media Independent Handover (MIH) is a standard that has been developed by IEEE 802.21 to allow the handover of IP sessions from one layer 2 access technology to other ones, to get a better mobility of end user devices. Figure 14 illustrates how the vertical handover architecture works. The mobile terminals have access to various networks like WLAN or 3G, which can have overlapping coverage areas. The User Agent has separate interfaces; each one receives its IP address from the corresponding wireless network. There are certain workarounds like:

- The mobility of the UA among various access networks is managed by the SBC.
- The mobility is controlled by the UA.

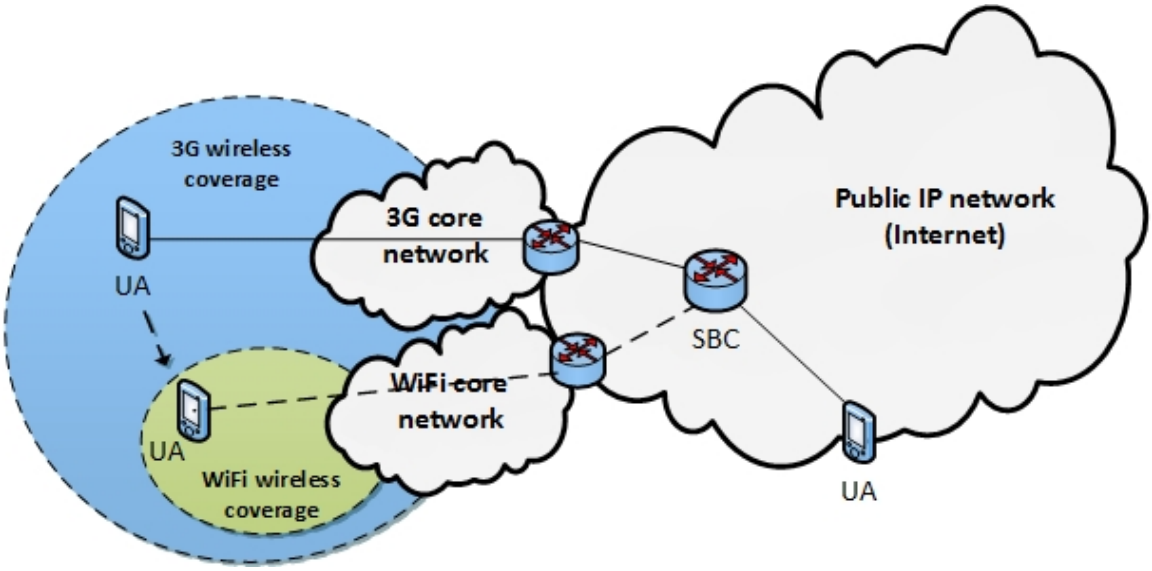


Figure 14: Vertical handover

2.6 Session border controllers

A session border controller (SBC) is a device commonly used in VoIP networks and mostly placed at the border among two or more networks. They are not the best of the solutions, as they tend to break end-to-end security and impact feature agreements.

SBCs usually work between two different service provider networks in a shared environment, or between an access network and a backbone network to deliver service to residential and/or enterprise customers.

SIP-based SBCs can handle both signaling and media. SBCs regularly modify some SIP headers and message bodies that proxies are not allowed to. An SBC

structure can be analyzed in Figure 15. It is logically associated to the local network, and is in charge of functions such as controlling and protecting access to the local network from the outer network. The SBC itself is configured and managed by the organization operating the inner network or by the service provider.

SBCs operate on both session layer 5 and network layer 3; and as such, they can process both the signaling messages and the media streams, while performing a communication session. They supply layer 5 controls and management in the network, which can not be managed by any router or firewall.

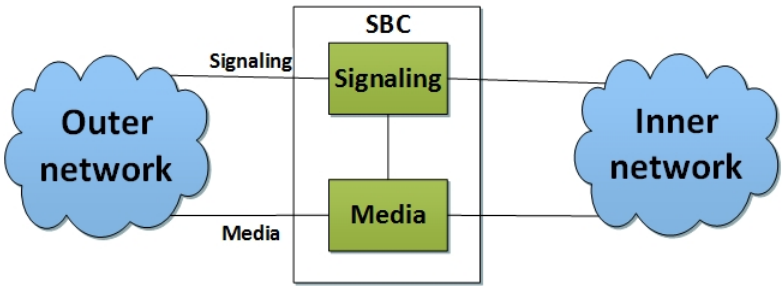


Figure 15: SBC architecture

A standard SIP call stream requires five messages to establish a connection and two messages to release it. Otherwise, in order for a call to be maintained, older SBC puts up with 100 RTP packets per second. This imbalance transforms a necessary SBC, into the bottleneck of all VoIP systems. It is impossible for older SBC's to choose a shorter path for media packets.

Figure 16 illustrates how a call works with an SBC scenario, it is divided in three parts, as the SBC is a termination point where the call ends and starts again towards the next network node.

2.6.1 NAT comparison

Figure 17 illustrates how the different packets are processed either with NAT or SBC. The number of sessions that NAT can process is much higher than with a SBC where all the packets are received and started again. NAT operates at layer 3 while SBC operates until layer 7.

Some peculiarities concerning NAT are described below:

- NATS allow external traffic only if it was initiated from the private LAN.
- VoIP calls that were initiated from the outside are disregarded by the NAT.
- The firewall does not understand some VoIP and media packets with embedded signaling.

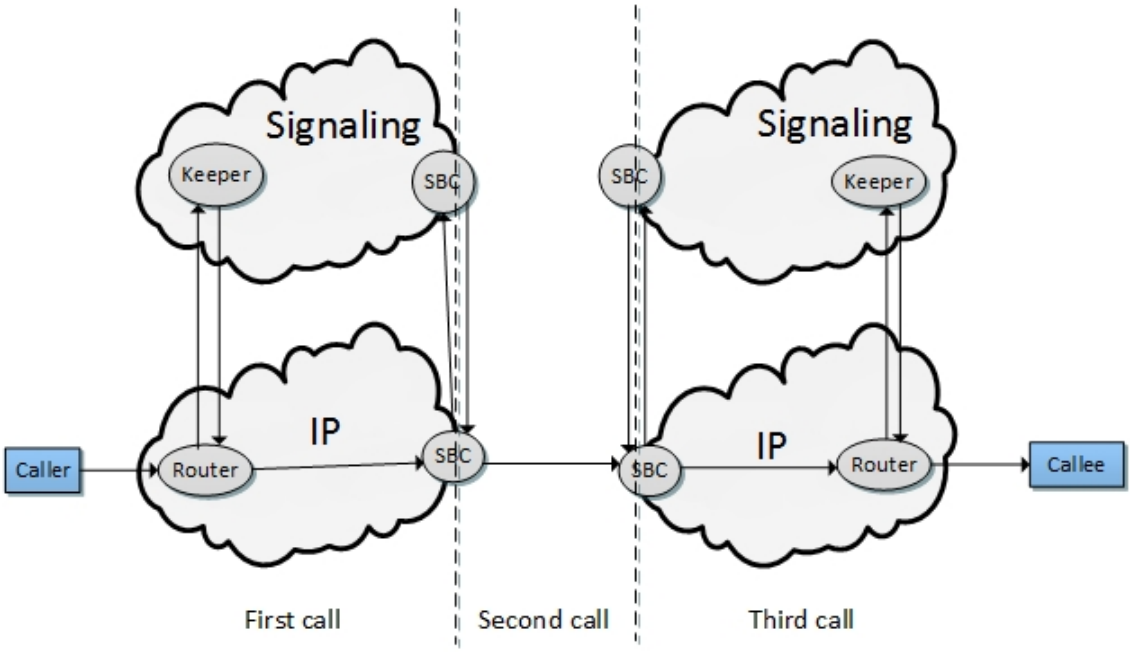


Figure 16: Call parts with SBCs

- VoIP generates various small packets and can easily overload traditional firewalls and NAT devices.

Application Level Gateway in NAT

Some of the application protocols are NAT-unfriendly. Such protocols are unable to traverse NATs without modification in the protocol messages.

ALGs (Application Level Gateway) are special translation agents that allow an application on a host in one address realm, to connect to its counterpart running on a different host and realm, without much opposition. ALG shares with NAT set up state, use NAT state information, modify application specific workload and any other action needed for the application to run.

Regardless of that, ALGs can survive without sharing any information with NAT, as only certain specifications may be needed. ALG make it easier for Applications to communicate with clients and servers, much as Proxies do. ALG's do not need any special protocols to communicate with clients as opposed to Proxies.

FTP ALG

Even while being one of the most popular applications on the Internet, FTP does not work with NAT. FTP NAT is then an integral part for most of its applications. Some service providers may even add some ALG's for customer support.

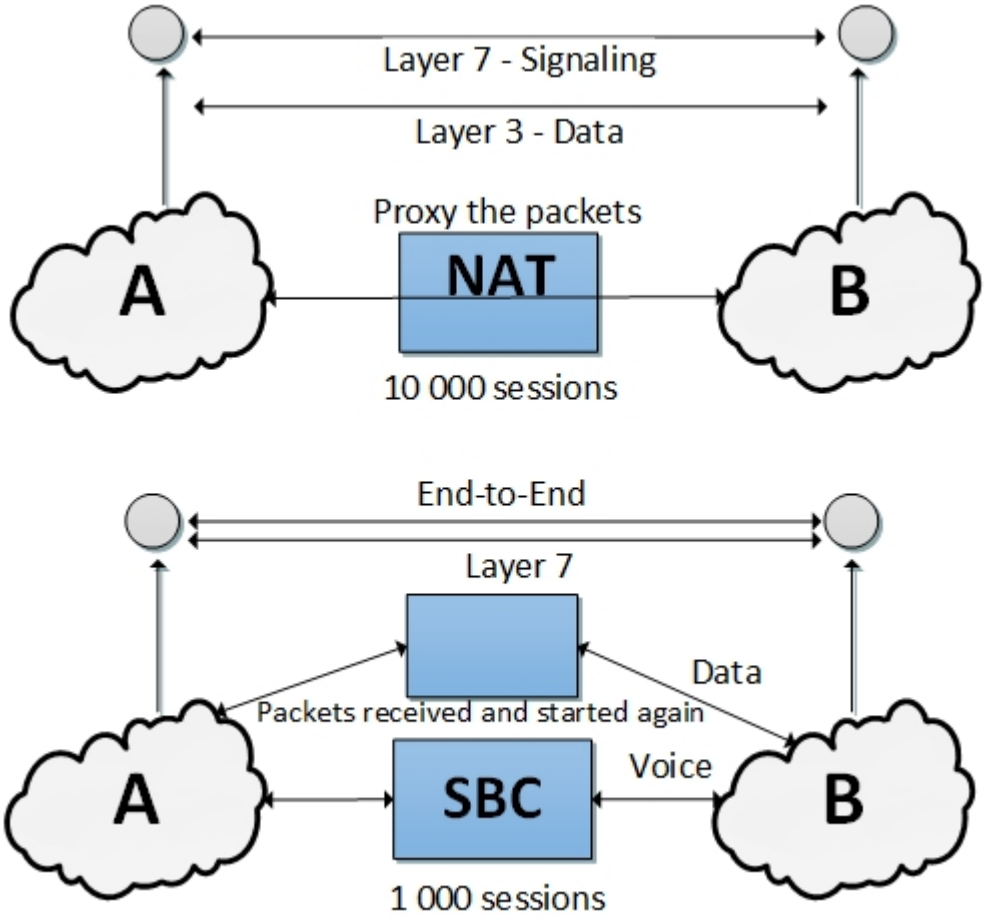


Figure 17: NAT vs. SBC

An FTP ALG is used to police and update the FTP control session payload so that information contained in the payload is pertinent to end nodes. ALG may also give some information to NAT in order to set up state for the FTP sessions.

2.7 Binary SIP

The most common language for humans to acknowledge or comprehend is the text message. SIP uses an encoded format using UTF-8 text encoding, but for networks, this means that the messages are large, need more resources and finally are less desirable for networks with bandwidth, delay or processing issues. As such, using SIP can sometimes result in delays, packets loss or exceed MTU size when running over WAN. Binary SIP encoding can reduce these problems as it helps to reduce the burden on the system. The difference between text protocol and binary protocol resides on what orientation each has, data structured or text strings. [32]

One of the most important resources in every system is the processing power. As any resource, it is finite and the service it can provide is limited by the workload it

can manage to process. Dealing with text based protocols requires a big amount of processing power, as it requires reading and parsing of every message. Text based protocols are easier for real people to debug, but use a lot of space. This affects no only the performance of the system, but also its cost. Memory is then consumed at a faster pace when processing text based SIP. The larger the server, the worst its performance is impacted with parsing and formatting.

When a series of servers are interconnected, each message can be converted from an object representation to a text representation (this is called formatting), or converted from text representation to an object representation (this is called parsing) many times as the message moves end-to-end through many different SIP proxy servers, or SIP B2BUA. This may amount to a large part of the processing resources.

Advantages of binary signaling:

- Latency: Smaller network traffic and quicker parsing reduces latency.
- Performance: Resources are less burdened by the faster operation. Better and faster parsing plus lower traffic translates into better use of the same resources as the same server can now handle more clients.
- Scalability: If each transaction is serviced faster, then the distributing or routing sessions becomes easier.

Disadvantages of binary signaling:

- No easy debugging: SIP can be human readable hence debugging is easier for a man managed network. But also software tools can be created to debug easily.
- Syncing client and server: Clients and server libraries need to be synchronized; without sync, parsing cannot be handled by the system. Protocol buffer ignores undisclosed extensions, so there is some freedom for an old client to connect to a newer server or the other way around.
- Firewalls-Existing equipment: New binary protocols cannot be utilized with existing ones. A SIP to binary SIP proxy is then needed to communicate.

3 Service Level Management

Generally speaking, the Service Level Management (SLM) refers to the different methods that a provider uses to handle quality of service that a given client requires. It serves the purpose of monitoring the Quality of Service (QoS) and the Key Performance Indicators (KPI) of a client. [29]

There are three main segments in which SLM is divided:

- Service Level Agreement (SLA).
- Quality of Service (QoS).
- Routing.

Before any services are provided to the final client, it is imperative to agree on the different services levels that may or may not be required by the specific customer. SLA is just the structure in which all those services are provided and thus supported by the service provider. The support for business processes has to be tailored to the needs a client has or may incur in the future in order to maintain the costs in check and so the prices charged can be held at the minimum average.

SLM makes it possible and easier to the service provider and its customer to maintain a constant check on the 3 segments described above. In that way, the technological support personnel can act faster; solve problems in the best and most effective and efficient manner, which in turn evolves into a better level of service to the customer. The business units can then have a better understanding of what do, when to do it and who is responsible for it. SLM encompass many activities, as it might be regarded as a checklist in order to ensure a better business practice and a reminder of the activities needed to develop a better project. These activities include:

- Identifying business requirements.
- Establishing the scope of services, timeliness, hours of operation, recovery aspects, and service performance.
- Translating business requirements into IT requirements and services.
- Developing and maintaining a service catalog, including costs for different levels of service performance.
- Develop gap analysis showing the business requirements and the available technological resources.
- Calculating the cost so that the services provided can fulfill the business needs and also the given budget.

- Creating, discussing and detailing different SLA with the customers, as to make sure their needs are fulfilled and that the provider can realistically deliver.
- Carry into effect the developed SLAs.
- Quantify the expected performance versus the actual one, and make adjustments accordingly.

The Figure 18 presents the interface between customer and service provider agreements by SLA and SLM specification.

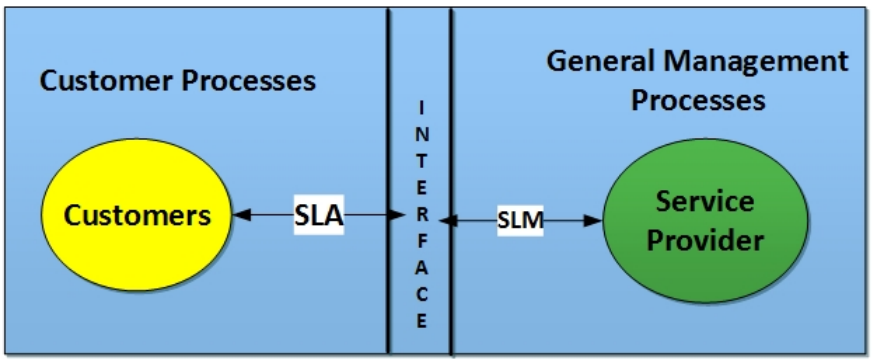


Figure 18: Interface between customer and service provider processes

3.1 SLA

The best way to describe a SLA is to compare it with a legal contract that ties the different parties involved to do certain activities and it must include:

- Defines the building pillars for the different areas of a project involving data, management, control and network service plans.
- Agree upon the setup and surveillance of end-to-end QoS inquiries among multiple carriers.
- Each SLA must be created in accordance to every client's needs. Furthermore, every agreement the parts come to, must be included in the SLA.
- The positions that all the parts involved within the business relationship will have, including the industrial partners (investors) and the actual working partners.
- Concerning the technical point of view of a SLA, it has to contain all the performance parameters needed for a project to be quantifiable and trustworthy like bandwidth, delay, jitter, packet losses as well as mobility management.

- The management component has to take into account, such characteristics as the network service set-up, its verification, testing, and recovery. This all has to comply with the OAM (Operations, Administration and Maintenance) phase.
- Intra and inter-domain QoS and end-to-end QoS guaranteed.
- On the monitoring performances side, Key Performance Indicator (KPI) are needed to measure the success of the project at large, and to see if the goals are being met.

During a VoIP service agreement, normally we speak of 3 different actors within the process:

- VoIP users: this normally refers to the final customers, which have to pay the service and are guaranteed a certain level of QoS.
- VoIP application providers: these are the ones responsible for balancing the necessary investment and operational expenses versus the revenue generated and also the ones providing the VoIP service to the user.
- The network service providers: These are the ones concerned with the transport service but they also have to create a good environment to have a good revenue-investment-operational expenses relation.

There is a relation between these three players as they all interact in many ways with each other. The VoIP users and application provider work in a sing session level Service Level Agreement (S-SLA). The VoIP providers and the Network service providers work in a network level Service Level Agreement (N-SLA).

3.2 QoS

As stated before, VoIP is a prospering and growing technology thanks to its great adaptability and flexibility. Because if the internet nature that this service has, it is important to have a special emphasis in the quality of service area, as the internet has a very different way of working than traditional communications.

This is where QoS comes into play, as it includes certain concepts that are imperative for quality like admission control or resource management. This gives the provider a good indicator as to what capacities or resources to increase depending in what requirements are needed from the customer.

This flexibility is what makes VoIP a little more cost efficient than the PSTN. The resources used to deliver these services are distributed among many players and thus also increasing the things that may go wrong. As such, there is a need for a certain level of quality, as voice communications require very little to no delay, low jitter and low loss. QoS is sometimes a necessity and it can be divided in two big areas: [7]

- Data plane: This plane involves packet classification, shaping, policing, buffer management, scheduling, loss recovery, and error concealment.
- Data control: This plane covers resource provisioning, traffic engineering, admission control, resource reservation and connection management.

QoS management architecture of VoIP can be partitioned into two planes: data plane and control plane. Mechanisms in data plane include packet classification, shaping, policing, buffer management, scheduling, loss recovery, and error concealment whereas mechanisms in control plane consist of resource provisioning, traffic engineering, admission control, resource reservation and connection management.

Unfortunately, classical SIP does not have a way to manage such resource prioritization. As a way to solve this inconvenience, the RFC 4412 RFC 4412 Communications Resource Priority for the Session Initiation Protocol, extend the SIP reach by adding an element that marks some requests inside the SIP. These are called: Resource-Priority header and Accept-resource-Priority header. They can be used by SIP users but also by PSTN gateways and terminals, to manage how the SIP request are handled and prioritized. There are certain situations where the Resource Priority can differ, like when the request is headed to a PSTN gateway, when the request is interrupted when it is a low-priority one, when there are multilevel priority domains in the PSTN and when there are higher priority request, as the SIP proxies and back-to-back user agents may overrule other requests. [4]

Two of the methods used to solve the problems with VoIP and its QoS requirements are the IntServ (RFC 1633) and DiffServ (RFC 2475). Both address QoS difficulties, and the second (DiffServ) is just answer to the first one is (IntServ) deficiencies.

IntServ comply with some of the most recent information formats like remote video, multimedia conferencing and visualization. It helps by delivering the end-to-end QoS that real-time applications require by managing network resources to provide QoS to specific user packet flows.

What IntServ does is to utilize the RSVP (Resource Reservation Protocol) to signal and reserve QoS for each stream in the network. With the RSVP, two kinds of service can be used: end-to-end reliability and guarantees bandwidth for certain kinds of traffic; and second, is a service that controls the load of packets given a moderate network traffic.

For IntServ to work, there has to be various functions with routers and switches:

- Admission Control: Determine if a new stream of packets can be given the requested QoS without affecting existing requests.
- Classification: Discern packets that need a particular level of QoS.

- Policing: When the situation requires it, packets have to be dropped in order to comply with the specified SLA.
- Queuing and Scheduling: Send packets according to those QoS requests that have been already been granted.

DiffServ is a modifiable end-to-end QoS mode that usually works within a domain. Its objective is to address certain difficulties with IntServ and RSVP:

- Scalability: Keep up with states by routers in high-speed networks is complicated due to the very large number of streams.
- Flexible Service Models: IntServ has only two classes of service models, but DiffServ offer many more.
- Simpler signaling (than RSVP): Some applications and end users may only want to specify a more qualitative form of service.

When employing DiffServ, a tiny bit-pattern in each packet is utilized to label it to receive a particular forwarding handling, or PHB (Per-Hop Behavior), at each network node. This is in accordance to the IPv4 (RFC 791) ToS (Terms of Service) octet or the IPv6 traffic class octet. Using the IPv4 ToS byte and PHB is the main characteristic of DiffServ:

- Packet Marking: The ToS byte is mainly redefined; six bits are now used to arrange packets. The six bits replace the three IP-precedence bits, and is called the Differentiated Services (DiffServ) Code Point (DSCP).
- PHB (Per-Hop Behaviors): When a group of data packets have the same DSCP value and they are crossing in a particular direction, it is then called a Behavior Aggregate (BA). It could be that some packets originated from different sources could have the same Behavior Aggregate. Per-Hop Behavior controls how any given packet is scheduled. Depending on the specific SLA or clauses agreed by the provider and client, it may affect scheduling, queuing, policing or shaping behavior of a node with an individual node.

DiffServ region may include:

- Static provisioning.
- Dynamic provisioning using RSVP.
- Bandwidth Brokers.

IntServ vs. DiffServ

When providing QoS on the Internet, it can at least be done through both DiffServ and IntServ. DiffServ differentiates the traffic while IntServ creates a virtual

circuit on the Internet using a special procedure reserving bandwidth.

IntServ works by remembering the state information contained in the routers. On the other side, DiffServ does not work by remembering any state in the network or the nodes. As the Internet can become very congested at certain times as it is a worldwide network, reserving and memorize all the state information can be a very complicated endeavor. IntServ is, therefore much more practical and useful for small private or corporate networks. As DiffServ does not have to deal with all that data memorizing it has more applicability for larger networks. To illustrate how the DiffServ mode works, the following figure shows how it operates. [5]

The Figure 19 shows the DiffServ QoS building blocks.

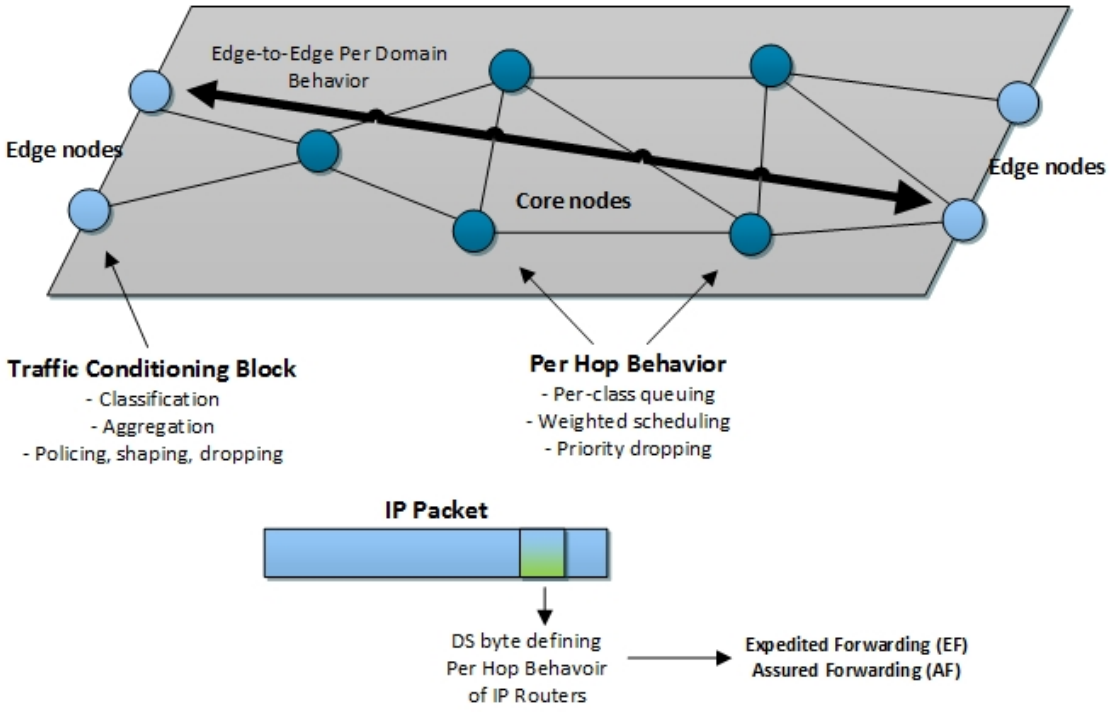


Figure 19: DiffServ QoS building blocks

QoS for emergency calls

Among some of the most helpful SIP characteristics, we can find that it can be utilized to support third party call control. This feature, allows us to give an emergency call or a high priority communication the necessary bandwidth to go through by denying certain other non-essential calls. This in turn, increases the possibility of the emergency calls to be successfully delivered.

In these situations, DiffServ is also useful, as it gives preferential treatment to marked IP packets. With the help of the Media Gateway (MGW), all the packets coming from a known IP address can then, be identified by the DiffServ. The core

network accepts or rejects packets depending on the label from the MGW. At the same time, the MGW is notified by the server, which packets have to be marked. Using that logic, all the emergency calls are marked and then recognized by the core network so that they can be given a best effort service. It can reject certain calls with less importance. In this sense, the MGW can be recognized as an admission controller, allowing packets to be marked for a better QoS. [24]

3.2.1 Call Admission Control (CAC)

It is widely accepted that end-to-end (E2E) delay is the main cause of QoS issues on the Internet. The International Telecommunication Union (ITU) states that the maximum acceptable delay for a E2E for a voice communication should not go over 150 ms. CAC can be defined as a choice made before any communication is made, like it used to happen with traditional PSTN networks. [25]

When we discuss CAC during a VoIP QoS situation, there has to be certain algorithms involved on the intermediate nodes when packets are being sent. CAC works with non traditional communications as it is based on when the given resources of a network can provide the necessary QoS detailed in the SLA. Within VoIP networks, there are three different groups of CAC: [14]

- Local CAC mechanism: Based on nodal information.
- Measurement Based CAC mechanism: Looks ahead into packet network in order to determine whether a new call is allowed.
- Resource-based CAC mechanism: Contains those that calculate resources needed/available, and those saving resources for the call.

A BB (Bandwidth Broker) serves as a tool for implementing bilateral SLA negotiations between close subnets, distributing bandwidth when it is requested or needed in coordination with the E2E CAC. For QoS assurance, there needs to be a BB that plays the roll of a gateway for any IP Subnet (ISP, carrier, enterprise). Utilizing two different kinds of algorithms for Call Admission Control, BB delivers signaling mechanisms for E2E decisions. These algorithms are: per flow E2E guaranteed delay services, and Class-based admission control in a core stateless network. When there is a new request, the Call Admission Control compares the bandwidth needed versus the available resources on the network. [12]

The Call Server (CS) is then in charge of accepting or rejecting calls or communications within the IP network, also managing the user profiles, while the SBC (Session Boarder Controller) concerns itself with administration of the resources. This collaboration between CS and SBC is possible thanks to the 'SIP Priority Header'.

In other words, CAC is responsible for the reception or rejection of new calls, all the while making sure there are the needed resources to carry the call. CAC

can become involved with tasks such as: limit the number of transmissions that can be supported by the network or the CPU, the user profiles, QoS of the links or outbound and inbound traffic loads. The CAC mechanism is divided between CS and SBC.

The role of CAC is to determine whether or not a new call can be admitted. Conventional CAC schemes take into account the inbound traffic load, outbound traffic load, QoS of the links, user profile, thresholds defined to limit the number of voice/video calls and CPU load. The policy based CAC mechanism has been split and performed at distinct locations (CS and SBC). Figure 20 shows this relation and cooperation.

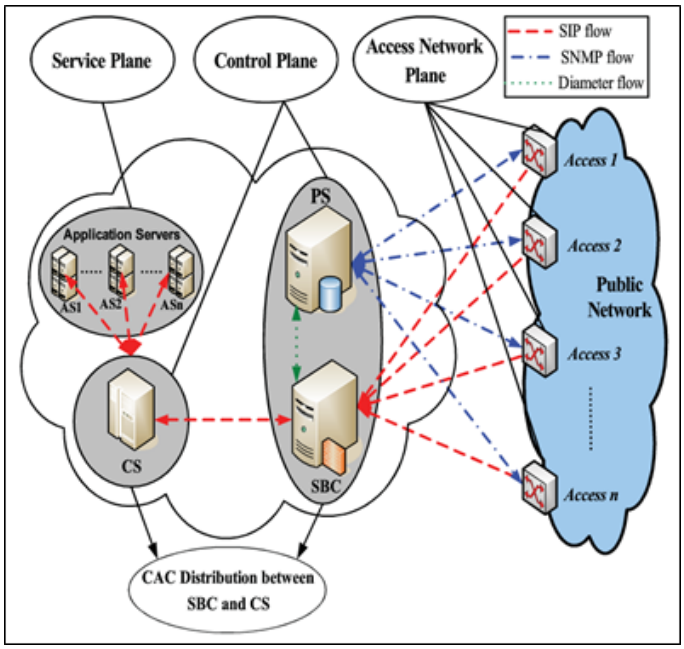


Figure 20: CAC

We can then summarize that the main components involved in CAC are:

- CS (Call Server): This feature can also support proxy, registration, redirection and location services.
- SBC (Session Boarder Control): Perimeter Defense (access control, topology hiding, DOS prevention and detection), features not supported at the end points (protocol interworking and media repair) and network management (traffic surveillance, shaping and QoS).
- PS (Policy Server): This has to be in accordance with the SLA. The PS provides the framework for any decision-making as it has all the statistical data and dynamic information. The policy system isolates the service, control and access planes.

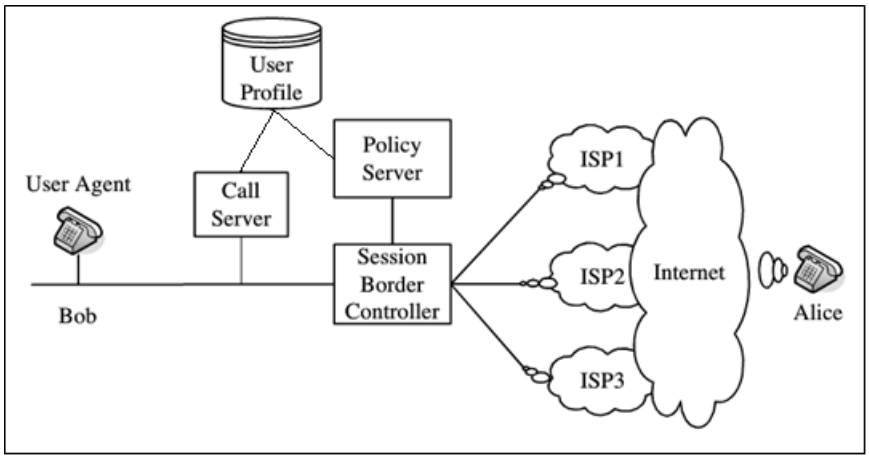


Figure 21: SIP CS, SBC and PS

Figure 21 shows components such as SIP CS, SBC, and Policy Server (PS).

Any CAC that takes into account resources of a network for QoS purposes has to deal with a SBC that interconnects any private and public network. It always needs to know and manage availability, costs, and quality of the transmissions. The CS shares the responsibility as it deals with any profile based CAC. Furthermore, PS comes into play when all the statistics (bandwidth, loss ration, call details records, etc.) are needed for a decision. [25]

Overprovisioning of communication resources is a method generally implemented to provide QoS in network backbones to avoid network control difficulties. Any node can make an admission control decision but it can also just only collect data for some other entity to use.

To intercept a call's signaling we use SIP proxies in order to improve the CAC. The most popular PBX (Private Branch Exchange) software, which is a system that manages calls between local and public lines, is Asterisk, developed by Digium. It ties up two different calls, one coming from the origin and the other one to the destination. Then, the SIP proxy controlling the local calls integrates with the PBX and the phones. Even calls that come through the PSTN can be controlled. This is called a SIP redirect server and it sends the call to another central office. [13]

UMTS mobile access

There are four different QoS classes for the UMTS mobile access according to the 3GPP (3rd Generation Partnership Project):

- Convensational (voice).
- Streaming (video).
- Interactive (web browsing).

- Background (emails).

With the guidance of the PDP-CA (Packet Data Protocol Context Activation Protocol), using also a predetermined GPRS; a MT (Mobile Terminal) decides for a determined QoS class for the UMTS network. Depending on what kind of application we are running, it can decide for any of the aforementioned classes. Particularly, when talking about voice applications, it uses conventional, but it can also utilise streaming for example for video streaming applications. [9]

In the UMTS access area of the network, resources are normally scarce and so the GGSN (Gateway GPRS Support Node) acts as a DiffServ router and a reservation protocol is needed for admission of some multimedia streams. 3GPP determines that a GGSN transmission has a dedicated SIP proxy, which is necessary for the communication between SIP call signaling and a PDP-CA resource reservation. The SIP then is in charge of the E2E CAC.

During the inbound phase, the DiffServ performs the monitoring of the streams based on the agreement with the Service Level Agreement. The SIP proxies then manages all the domains they control and make the decisions and what is to happen with any information that passes through, in accordance with the SLA and SIP. [33]

Figure 22 shows how the UMTS IP interworking would be carried out.

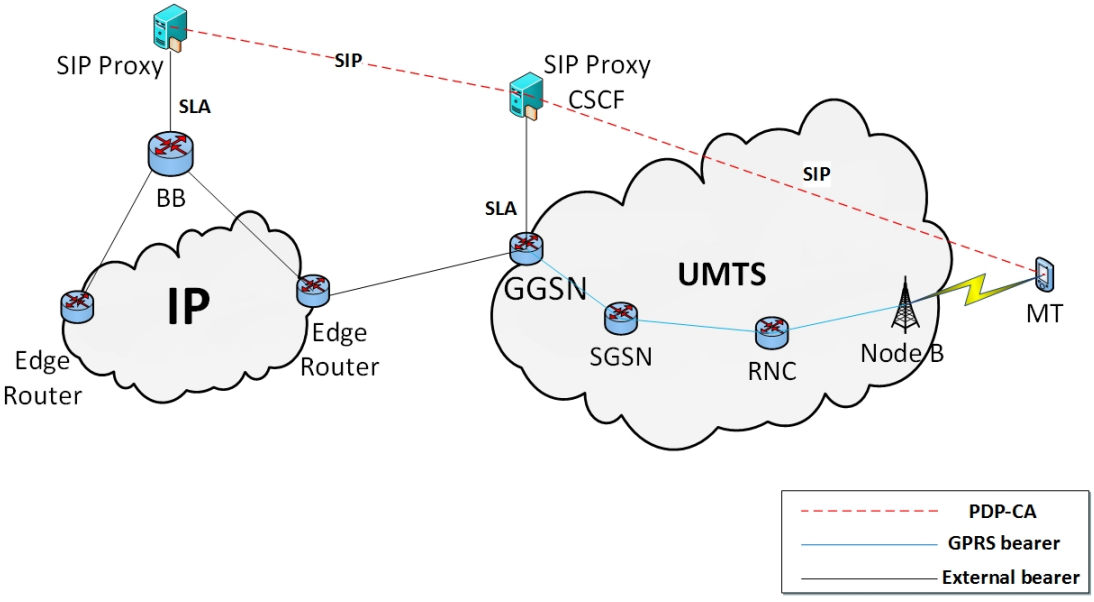


Figure 22: UMTS IP interworking

3.2.2 Bandwidth Broker

A Bandwidth Broker (BB) is in charge of executing QoS control and management activities such as admission control, resource reservation and provisioning for an entire network domain. As stated before, it works in coordination with the DiffServ model and controls the traffic within a network.

Some labels are put on certain data packets from certain customers and monitored to comply with the SLA, as the core network is not capable of discerning between traffic from different clients. The BB is composed at least by two separate database components: [16]

- The Network Resource Database (NRDB): Contains information concerning resources in the network like IP-level topology, provisioned and available link capacity, and paths between nodes.
- The Policy Database: Accommodates network-wide service statistic information for each client.

As stated earlier, certain QoS guidelines have to be followed to comply with the SLA. Every time a stream of packets or data is received, some bandwidth has to be reserved; in order to do this, the edge router has to request a reservation from the BB.

After that, it is up to the BB to admit or reject the stream. The BB performs a test that determines the QoS state and also knows if there are enough resources to complete the request, in this case, bandwidth. If there is enough bandwidth to accept the stream, then the QoS state is updated, and with it, the information itself. This makes a new reservation for the incoming stream. If, on the other hand, there is not enough bandwidth, the reservation is rejected and no QoS state is changed.

Whatever the BB decision, the edge router then is informed what is to be done. In the case of a reservation tear down request, the BB updates the link, information and path state database, so to account for the leaving stream.

Another task under the surveillance of BB, is the vertical integration of the various elements involved in the requests: data, control and management. The BB governs the domain's network resources and so it is able to integrate them seamlessly.

Unfortunately, when a BB is used for admission control of VoIP calls, some extra call setup delay will incur. [17]

The main objective when utilizing the BB architecture is to manage several services inside the network domain with the ultimate purpose of delivering a reliable QoS. All the domain variables: information of the network resources, domain topology, service policies and SLS (Service Level Specifications) are concentrated and this results in fewer complications for the network. This, in turn, allows the network core

to perform control activities or changing the state of the information.

The process starts with the signaling message arriving to the BB with a certain flow profile and QoS requirements. After authenticating the request, the BB makes the corresponding decision depending on the available resources, the domain service policies and the SLS where the BB interacts with the nodes at the data plane. [15]

Figure 23 shows how BBs place at the control plane interact with edge nodes at the data plane.

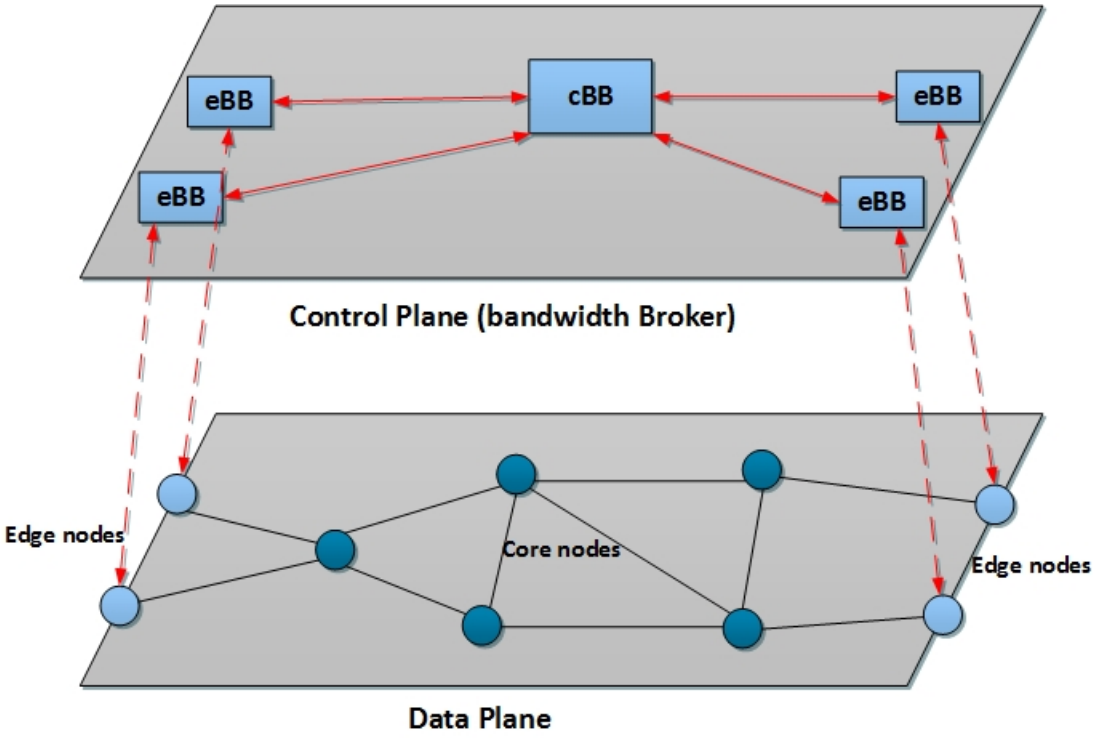


Figure 23: Bandwidth Broker

3.2.3 Multi-Level Precedence and Preemption

Multi-Level Precedence and Preemption (MLPP) contributes with the prioritized call managing service. It is made of 2 elements:

- Precedence: assigns priority level to a given transmission.
- Preemption: Finds and allocates resources for the transmission.

The service provider, depending on the customer's needs, agrees upon all the parameters concerning what the priority or precedence level is for a specific customer, at the very beginning. All this prioritization and MLLPP is handled by the MGC (Media Gateway Controller):

- Preemption: Certain low priority calls may be interrupted or dropped to make way for a higher priority one. Different calls need different resources like bandwidth, so a single higher priority call may interrupt many lower priority ones.

- Priority queuing: Bandwidth resources are logically finite, and due to that reason, a queue is then created in the order specified by priority of a given call. It might also be the case, that the queuing has a different order, for example, unless otherwise specified, the calls are served on a first come first served basis, where the call that waited the most is the first one to be served, of course, within the capacity of calls for every queue. The MGC checks if there are available resources and then give preference to the first non -empty queue respecting the priority according the service agreement. Additionally, there might be fixed times for a call to wait in queue, and after the specified period the call will be dropped or fail. Generally speaking, the dropped calls might always be the lower priority ones, and since the call was never really established, it fulfills the requirements for the preemption. [3]

With circuit-switched networks, or traditional ones, there is a service called Precedence-Based Assured Service (PBAS). It is used to maintain a certain level of QoS and reliability to critical communications, regardless of the allocated resources at the time. This rules, are imbued into the system to allow for the minimal accepted level of service. These circuit systems posses the following characteristics: [5]

- Precedence and preemption.

- Five normal levels of precedence.

- The service provider sets the maximum precedence level of a customer at the subscription time.

- Lower precedence MLPP calls can be dropped by a higher precedence call.

- The MLPP service is defined for a domain of a network, which can be both, a subnetwork or a complete one.

- A non-MLPP domain that transmits MLPP calls has to conserve the MLPP marking within the domain.

Extending the MLPP functionality: [6]

- The legacy voice network and the VoIP network share similar call control architectures.

- With standard call signaling protocols like SS7 and SIP, call control has to be introduced. To implement certain features that are similar to MLPP, Assured Services SIP (AS-SIP) have extensions that work with precedence-based designations.

- In normal traditional transport networks, legacy voice works with TDM (Time Division Multiplexing). In other words, it has dedicated resources for the transmission. VoIP works differently, as it shares all the various resources of the network to make a more efficient use of the infrastructure.

- Due to the shared nature of VoIP networks they need many control and managements systems that ensure some QoS.

3.3 Routing

Thanks to IP technologies, the transmission of packets, and in this case VoIP can benefit form many kinds of call setup, voice transmission and resource management systems.

Usually, delay and jitter were the biggest problems for VoIP technology and data packet transmission, but lately, the call setup and routing has moved up in the priorities agenda because it can also affect the routing options of a given call. Call routing is involved in the following three elements: [18]

- There is a possibility that because of insufficient resources, a call might not be blocked along the signaling path.

- The QoS can have errors with the prioritization of the data packets during traffic and information transmission may not be correct.

- Post dialing delay my occur, which means too much time passes to hear a ringtone or call setup delay.

Many of the network protocols already described in this thesis help to ameliorate the problems related with quality as it has become easier to manage traffic. Providers of VoIP, with the goal of increasing quality service, have relied on some of these protocols to create modern call routing services. SIP belongs to these protocols, just as ENUM (Electronic Number Mapping System), which helps with the following characteristics:

- Converts conventional telephone numbers to SIP addresses.

- Its DNS-based architecture enables answering ENUM requests from end-user devises and signaling servers.

These network tools allow for communication and interoperation between the Internet and the traditional circuit-switched infrastructure to make and receive voice calls. A common routing policy for this exact case could be how the service providers choose from the many gateways to complete a transmission.

ENUM and SIP protocols in accordance with routing policies can be used for circuit-switched networks and their routing policies. In the case of long distance

or international calls, telephony gateways can bypass toll charges. This can help many local providers, but these providers also have to have agreements with other providers in different locations for a call to be accepted. It could also mean that some providers may want to standardize their routing policies, in order to have a similar call setup delay and reduce the time users have to wait.

This business model, involves many different parties and providers such as gateway operators and aggregators, which are the Internet Telephony Service Providers (ITSP) or Inter-exchange Carriers (IXC). To fulfill customer demands and to establish a given connection, many gateways are involved and so, the service relies not only on one entity. [19]

3.3.1 Dynamic routing

Dynamic routing refers to a technique used to provide the best routing possible for a call according to real-time network alterations. There are two classifications of dynamic routing protocol:

IGP (Interior Gateway Protocol): This protocol is used inside a local single administrative domain and can be based on two alternative link-state routing protocols:

- OSPF (Open Shortest Path First): This points to a link-state routing protocol within a particular network system. IS-IS (Intermediate System to Intermediate System) is another protocol that also uses SPF more used as it makes the data transmission more efficient.

EGP (Exterior Gateway Protocol): This protocol is used to make different autonomous systems compatible. In practice, the exterior routing protocol in use is the BGP (Border Gateway Protocol).

There are certain advantages that dynamic routing has over static routing. Static routing means basically to manage the routes of a given call manually. Dynamic routing, thanks to the protocols above, has more scalability and adaptability. Being able to adapt more rapidly to changes in the network means it can react better to possible errors or failures of the system.At the same time a dynamic routing protocol makes routing changes automatically, leaving less control to the operator; this can be seen as a detriment in the operator side to achieve 'Carrier Grade' service.

Dynamic routing protocols

As stated before, the protocols used for link-state routing are OSPF and IS-IS. They both depend on the Dijkstra algorithm to work and find the shortest route within a certain network. [31]

Compared to OSPF, IS-IS helps managing the router convergence time, does not use so many bandwidth requirements, and is a more reliable and scalable to larger Link-State Databases (LSDB). The purpose of both protocols is to interconnect a specified network with other routers. These protocols are widely utilized, as they are the best suited for most systems and configurations; they comply with most security, bandwidth, memory, time and CPU requirements. IS-IS has become the standard link-state protocol with most Communication Service Providers, as it is the more flexible of the two.

Routers are needed in most networks for controlling and redirecting data among other routers and networks. IS-IS is used in most efficient networks, but especially for large networks as the workloads cannot be managed optimally with static routing. As such, choosing the appropriate routing protocols is critical to have a functional network. Close ties among providers have a great impact as traffic and speeds may vary among regions.

When a network uses dynamic routing protocols, routers can connect and unveil their available information paths to other routers faster. All the configuration changes can be dealt with while avoiding certain system failures.

Link state protocol vs. distance vector protocol

Within the Interior Gateway Protocol, we have 2 kinds of protocols that manage the routing paths of the information transmitted: Link-State Protocols and Distance Vector Protocols. Both fulfill the same objective, but one does it based on a distant metric and the other one based on an interface, in other words, it selects a path using the state information of each link.

Distance vector protocols sends information on the best path they can find to a destination, while Link-state protocols work with both link and node information from their networks.

Link state protocol

These kinds of protocols need to keep a complete awareness of their network to work properly. Using Link State Advertisements (LSA) the network notifies the routers of link-state changes and then, based on a link-state database, the router calculates the optimal path for a given transmission. This process is done based on the Dijkstra algorithm and the results are stored in the route table. To create a link-state environment, the first step is neighbor discovery, and sending a 'Hello' package to start it. The connection is established and LSA's start flooding. There are many ways to make this process more efficient and reliable, like using unicast and multicast addresses, checksums, and positive acknowledgements. There are 2 kinds of Link-state protocols:

Open Shortest Path First (OSPF)

The OSPF is defined in RFC 2328. It is one of the possble alternatives to the Interior Gateway Routing protocols. A router works by identifying itself with all the other neighboring routers and synchronizing databases. Routers keep communication regularly with each other by means of a link state advertisements that is stored to make up the topology map of every router interconnected. The routers decide the best possible path for a call using the 'shortest path tree' to the closest router available and communication is established.

Intermediate System Intermediate System (IS-IS)

The IS-IS protocol is defined in RFC 1142. It optimizes the Interior Gateway Routing mechanism. When this protocol sends its local state information (eg. usable interfaces or reachable neighbors) to other routers, it also shares the cost of utilizing said interface with a Link-State Payload Data Unit message. With this information, the routers create a topology of the Autonomous System (AS) and each one of them designs their own best path using the Dijkstra algorithm.

The routing table stores the information regarding all the destinations, associated with the other hop IP addresses and outgoing interface. This process is dynamic as the routers redo this process every time there are topology changes. It provides assistance for multiple paths with the same cost, multi-level hierarchy so it may hide itself from routers that belong to other areas (only trusted routers can be part of the exchange), improved routing protection and minimizes the protocol traffic.

Distance vector protocol

Generally speaking, routing involves two elemental tasks: Identifying an optimal route and the forwarding of the data packets within the network. The most complex of the two tasks is finding the best path. One protocol that works to determine such path is the Border Gateway Protocol (BGP).

BGP works with interdomain routing in TCP/IP networks. BGP is based on the Exterior Gateway Protocols (EGP), and is in charge of routing across multiple autonomous systems and communicates/connects with other BGP systems.

BGP was thought in principle, to improve on and replace its predecessor, the Exterior Gateway Protocol (EGP). BGP resolves important problems with EGP and scales to Internet growth and traffic in a better way.

Boarder Gateway Protocol

This standard is defined in RFC 1771, which describes BGPv4, the current actual version of BGP. Similar to some other routing protocols, BGP deals with rout-

ing tables, transmits routing updates, bases routing decisions and collects routing statistics. The most important aspect of a BGP system is to exchange network-reachability information, with information about the list of autonomous system paths, as well as with other BGP systems. This information must then be used to develop a graph of autonomous systems and their connectivity so correct routing decisions can be made.

BGP bases its analysis on the routing metrics to determine the best route to a given network. This metric contains a number that specifies the degree of preference of a determined link. The network administrator typically assigns the BGP metric to each link. These numbers can be based on many characteristics, the number of autonomous systems through which the path passes, stability, speed, delay, or cost. Unfortunately, BGP suffers from severe and frequent routing changes and slow convergence.

TRIP

To establish VoIP communication, SIP works with a text message structure within a network. In order to provide next-hop routing information for calls, it utilizes its Location Server (LS) function. Unfortunately, LS is not useful incapable of working with dynamic routing, and thus it is not useful to the creation of the Telephony Routing over IP (TRIP). This is defined in the RFC 3219.

TRIP is another routing protocol that works in parallel with the IP network. It collaborates, by constructing the necessary routing tables for the proxy it supports. This information will ease the proxy's duty to make session request decisions and find a suitable gateway for the given data. In other words, TRIP finds the best gateway from a VoIP network to a traditional PSTN.

TRIP send its requests through a Transmission Control Protocol, as thanks to that, TRIP does not need to retransmit, acknowledge or sequence the communication. Hence, TRIP is independent of any other signaling protocol and can work with any other routing protocol.

Mainly, TRIP is a protocol for advertising the reachability of telephony destinations among servers, and also to distribute the routing information to other routers. For this purpose, the BGP-4 is put in use to share the information among administrative domains. TRIP follows the steps of BGP-4 (the generally accepted standard for internet routing) as they have similar tasks. The problem with TRIP and BGP-4 is that the routing decision is taken under blocking uncertainty, as they do not share dynamic gateway state information among VoIP providers.

Routing for VoIP

When a VoIP call is established, the routing deciding the path that the data or

media flow must traverse, is always decided at the very beginning of the transmission. Generally, every connection is established almost instantaneously, but sometimes a certain interaction is needed, like the destination-required time to pick up the phone. The measurements to decide on the best possible path are made during this period of time, assuring the calculations have some space to finish. If there is not enough time to make those calculations, it may happen that the path decision is not the best one. This could happen for example when automated robots answer certain calls without any delay.

Routing VoIP calls in IMS

Considering MGCF (Media Gateway Control Function) node as the one that can work with both circuit-switched and packet-switched address formats and provides interworking between CS and IMS/PS domain, there could be routing problems for VoIP calls in bigger IMS networks.

Certainly, networks are growing, and that means that better and more reliable Media Gateways have to be created. The problem can happen during a session set-up, when we are attempting to find the appropriate MGCF node terminating the call to the end-user in the CS network.

The destination addresses cannot be translated until the call enters the MGCF node. If the MGCF was chosen poorly, another MGCF node has to be accessed and the whole process repeated. This phenomenon is called Cranckback in CS routing. This obviously means redundant routing, and waste of resources, which in turn, increases unnecessary traffic load and delays. An appropriate solution could be to create a function, which would choose the appropriate MGCF node for routing the call across the technology boundary and to the destination.

Additionally, if all requirements concerning the IMS architecture are to be met and the routing problem solved in the most efficient way, a new function node between CSCF and MGCF nodes should be added - Breakout Gateway Control Function (BGCF).

4 Security

All VoIP technologies make use of a series of different protocols. Among these protocols, we encounter the signaling protocols such as SIP, but also data control and transfer protocols as TCP, RTP, UDP, and IP.

Generally speaking, telephone communications systems through IP, both public and private, revolves around the transmission of data packets, as opposed to circuit-switching in more conventional systems. These data packets are then decoded and reassembled on the receiving end.

There are three important characteristics when describing security in communication systems: [38]

- Confidentiality: This refers to the need for certain restrictions on which person or user can access the data packets, and thus, the information contained in the communication. Therefore, it handles different types of users and discern between authorized and non-authorized users. This aspect has a special importance as it also deals with the interception and illegal use of the information. It is probably the most important aspect view from an enterprise approach.

- Integrity: This aspect is defined in the absence of authorization what the system does and conserves certain information. For example: in case of changes, deletes the contents of the voice or data, passwords, configuration, and other stored information in the system.

- Availability: This characteristic has to do with the accessibility that an authorized user has to the system at any given time. It also refers to the availability of all the resources of said system as for example: storage, transmission or to be able to reach the user.

As stated previously, one of the most, if not the most important aspect of IP networks, is that of transfer of information. Authentication of an authorized user is crucial if a network is to be secure.

Certainly, many problems concerning QoS support, have to do with the IP network layer, but it is crucial not to disregard the security issues dealing with the control and service architecture and the many signaling protocols.

Unfortunately, security on VoIP systems cannot be addressed with a single solution. In order to maintain an optimal level of security throughout the SIP protocols, we have to resort to at least 2 different aspects and protect the signaling separately from the media. Figure 24 illustrates this simple security division and its branches.

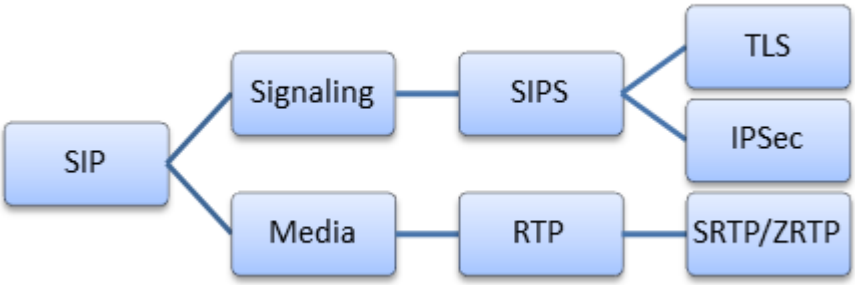


Figure 24: Security protocols

4.1 Protocols

4.1.1 SIPS

When sending SIP messages over a TLS (Transport Layer Security)-encrypted channel it is necessary to resort to SIP Secure Standard (RFC 3261). SIPS is used to make sure that any SIP taking place in a TLS is enabled among a pair of hops. This validates and secures the transmission and provides a secure connection from endpoint to endpoint.

The best way to ensure that a VoIP call will be secure is through the RFC 3261 standard, which names the SIPS URI (Uniform Resource Identifier), assuring that all the traffic within any given call using TLS will be secure from the user initiating the call and all the way to the receiving user.

This security mechanism works in a similar way that the HTTPS as it encrypts and then authenticates any given transmission on TLS.

4.1.2 SRTP

Just as the SIPS, SRTP is a security standard added to the RTP. It is collected in the RFC 3711 as published by the IETF (Internet Engineering Task Force).

This standard provides the necessary security in terms of authentication and confidentiality needed by the RTP. It works by means of a secret key that serves to an AES (Advance Encryption Standard) algorithm, which protects the transmission.

SRTP minimizes the number of key pairs that must be shared between the two nodes communicating. It is designed to add low overhead to the packet size.

4.1.3 ZRTP

The unique addition made by the ZRTP is the way in which the keys are used in an SRTP change. ZRTP is especially useful as the keys created are unique for every

session, so it is very helpful against authentication attacks. Thanks to these properties, ZRTP has better standards of confidentiality than normal SRTP, as no keys are available to any server prior to the transmission.

Moreover, ZRTP does not rely on SIP signaling for the key management, the secret shared between both endpoints is the one that generates the different keys for encryption. This protocol is defined in the RFC 6189 of the IETF.

4.2 SIP Security mechanisms

SIP works with many standard mechanisms that ensure the security for the information. The two issues that have to be regarded as important while securing the SIP header and body information are:

- Privacy: The SIP has to have all the information concerning the users and networks secure.
- Identification: The SIP sessions have to be created in a way that allows only for an authorized user to interact. This includes all the necessary security to prevent fake identities to access the information.

There are also at least two ways or mechanisms in which SIP can be secured. They are classified according to the form in which they interact with the system:

- End-to-end mechanisms: This method utilizes some features contained within the SIP. The SIP includes means to authenticate and encrypt the message. Both the caller and the end destination SIP are involved.
- Hop-by-hop mechanisms: This method relies on network level or transport level security. It is not contained within the SIP itself and works by securing the transmission between SIP nodes in the path of signaling messages.

There is a need for hop-by-hop security as there exist some elements in between the communication that may affect or change the SIP process. There are intermediate SIP actors that may also modify the message, and this is where hop-by-hop security comes into play.

The most used mechanisms to provide security for SIP are:

- Authentication: This can be achieved by both end-to-end and hop-to-hop mechanisms. Any secure SIP communication has to authenticate the sender and also has to make sure that the information contained in the message was not modified along the way. As stated before, SIP counts already with some security and authentication mechanisms like proxy-authenticate, proxy-authorization that serves as a digital way to ensure the identity of the sender and the message, but there also is need a for some hop-by-hop authentication too. TLS or IPsec (Internet Protocol Security) can be used along the way to authenticate and encrypt the packets of information.

- Data encryption: As the name implies, this mechanism lets only the sender and the destination user access the information contained in the communication. Data encryption, as the authentication mechanism can be achieved by means of both hob-by-hop and end-to-end. There are many encryption algorithms used to secure a message (header and body) but the most common are DES (Data Encryption Standard) and AES (Advances Encryption Standards). The common end-to-end transmission is encrypted by S/MIME (Secure/Multipurpose Internet Mail Extensions), which ensures the identity of the sender as well as encrypts the messages or emails.

SIP contains end-to-end encryption but as it was defined earlier, it only can manage security for transmissions where no intermediation between the sender and receiver is happening. This only targets messages or information that will not be going through any other proxy server. When the message goes through intermediaries, hop-by-hop security is needed and TLS and IPsec are then used to encrypt the message and prevent any unwanted party to access the information.

MIME (Multipurpose Internet Mail Standards) are a group of characteristics that enrich in many ways an email message. Every SIP message includes MIME bodies and within these characteristics, the standards include integrity and confidentiality. S/MIME can encrypt and encapsulate SIP messages so as to provide a minimal degree of security.

One of the drawbacks of S/MIME is that messages can become very large and that it does not provide an environment in which there might be a public key exchange. But it only speaks of one of the security dilemmas, the more secure, the more trouble or inconvenience for the end-user.

SIP identity

Another possibly minor inconvenience regarding SIP security standards is that of the SIP Identity. SIP provides a way to exchange public keys with the PKI (Public Key Infrastructure). This mechanism provides security and integrity of a call or message by means of diverse public keys for verification purposes.

The problem with PKI and SIP Identities is that a Public Key Infrastructure is needed for each user, and the maintenance, management and operation of a user-level PKI signifies many costs, time and also the digital approval from a CA (Certificate Authority) with its periodical checkups.

4.2.1 IPsec/TLS

These are security mechanisms that apply to the hop-by-hop, which means, they work outside of SIP. The way in which SIP works is completely independent on how IPsec or TLS work.

IPsec works with network node identity and for that reason, it can be used between different SIP users that might have a preconfigured and static security association. This mainly refers to packets of information between two servers that have a lasting communication relationship.

TLS provides transport-layer security over connection-oriented protocols (TCP) and it is suited to architectures in which hop-by-hop security is required between hosts with a more dynamic security association.

When we use the hop-by-hop mechanism to provide security to a SIP request to a proxy server, we could make use of TLS or IPSec but it will also need end-to-end security provided by the SIP itself.

The SIP specification includes a way to specify that a resource (a server or user) should be reached securely using TLS. In particular, the address of a user is normally defined in SIP using a SIP URI (Uniform Resource Identifier). If a user address is expressed using a new type of URI, a SIP Secure (SIPS) URI, it means that the use of TLS is requested.

When TLS is used to provide security for a transmission of packets of data, the problem of it not running on UDP (User Datagram Protocol) may prove to be an inconvenience. It might be necessary to utilize long-lived TLS over TCP connections. TLS can only provide security and authentication to SIP communications through the hop-by-hop mechanism.

4.2.2 Security agreement mechanism

In a communication initiated by the client, the SIP agent includes in the first request sent to the next hop entity the list of its supported security mechanisms. The other party (the server side) responds with a list of its own security mechanisms and parameters. The client then selects the highest-preference common security mechanism, turns on the selected security and again contacts the server using the new security mechanism.

4.3 Security Architecture

The Security architecture concerning the VoIP communications can be divided into three different security domains:

- Trusted domain.
- Trusted-but-vulnerable domain.
- Untrusted domain.

Figure 25 shows the different domains with their network elements.

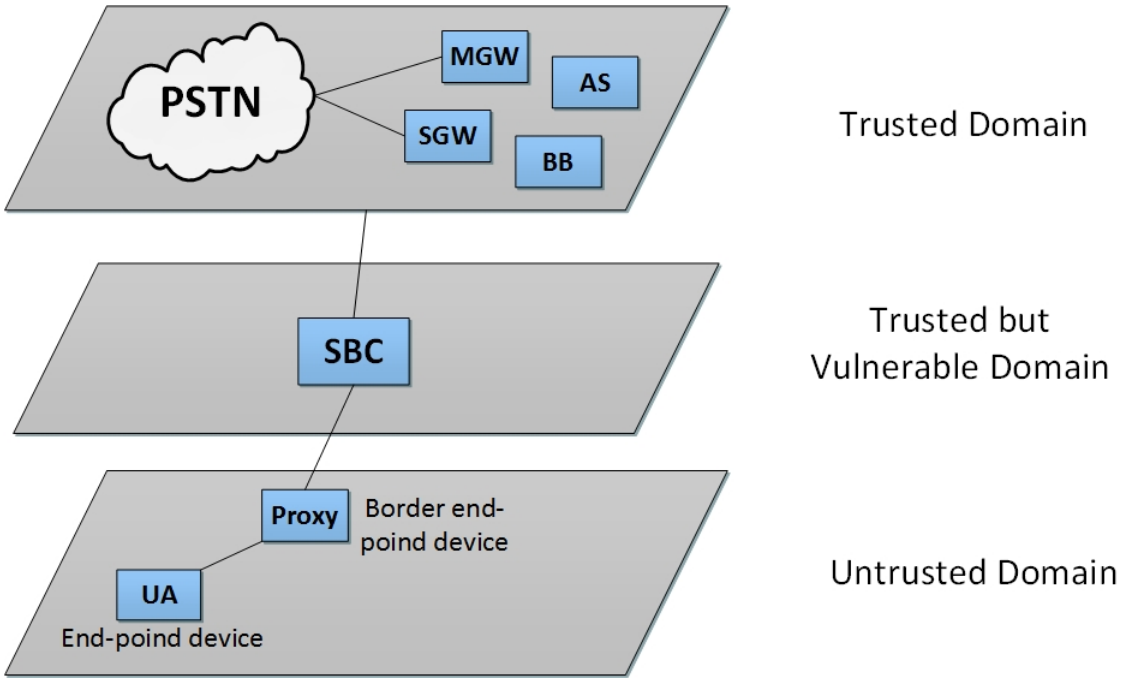


Figure 25: Security domains

4.3.1 Trusted Domain

A Trusted Domain includes any server that helps to deliver a VoIP communication. These contain but are not restricted to the AS (Application Servers), call control elements, BB (Bandwidth Broker) and CAC (Call Admission Control).

Moreover, the Trusted Domain also deals with various gateways like MGW (Media Gate Way), SGW (Signaling Gate Way), MGWC (Media Gate Way Controller) which allow communications among many transport protocols on the more conventional PSTN (Public Switched Telephone Network).

As the name implies, a trusted domain is provided by the Service Provider or is an Enterprise Network. This comes with an increased physical security, as the geographical place where all the services are located is the same location. This means that the danger of intermediaries to exploit any message insecurity are reduced greatly, as the Trusted Domain never has any direct contact with any Untrusted Domain.

4.3.2 Trusted-but-Vulnerable Domain

The difference between a Trusted Domain and a Trusted-but-Vulnerable Domain is that it contains a SBC (Session Border Control), which controls the signaling of a VoIP call. There is contact with the border or outside the local secure network of the Trusted Domain as it can also communicate and interact with Untrusted Domains. Still, there is a level of security at hand, as it protects the infrastructure elements

from unusual occurrences like an increased volume of requests, badly formatted requests or excessive traffic.

Generally speaking, the Trusted-but-Vulnerable Domains are also under the jurisdiction of the Service Providers or the Enterprise Networks, but have interactions with the rest of the world or third party networks.

4.3.3 Untrusted Domain

The Untrusted Domain is comprised of any devise that is not under the Service Provider control. It includes for example: SIP phones, IP-PBXs, clients running 'softphone' applications and border elements that are not part of the Trusted Domain.

Untrusted Domains can only communicate with Trusted but-Vulnerable Domains through SBC.

4.3.4 Signaling Security Architecture

Signaling within the Trusted and Trusted-but-Vulnerable Domains

The safety of a Trusted Domain remains on its local and closed environment. Every communication of VoIP is contained inside this private and secure network that has no contact with any border element. Also, every transmission that occurs inside the Trusted Domain is authenticated via a unique certificate. A Trusted Domain provides its users with a secure communication and prevents any third party or outside intermediary to interfere or access in any way with the information contained in the message.

Signaling within the Trusted and Trusted-but-Vulnerable Domains

The contact point between Trusted Domains and Untrusted Domains is the Trusted-but-Vulnerable Domains. The contact point is a SBC that makes it possible for end-point devises and proxies, to communicate with Trusted Domains. SBC redirects the signaling from Untrusted to Trusted Domains.

SBC acts as door guard, as it has the power to accept or reject calls depending on the authentication of the sender. It also recognizes different service level agreements and certain capabilities of end-point users so as to have a better image of who can communicate with a Trusted Domain.

The SBC acting as a security guard that recognizes certain kinds of communications uses the service level agreement as the means to discern between different customers and authenticates accordingly. The SBC also rejects unauthorized packets of information based on the source rather than the individual customer.

Additionally, every communication must be encrypted using the already mentioned TLS or IPsec, as the signaling between nodes has to be authenticated with the verifiable certificates.

4.3.5 Media Security Architecture

SBC Media Security

The way a SP can control its media security is via the same SBC, as it can filter not only the signaling but also the media transferred.

Any media that comes from the end-point device to the SBC gives the Service Provider control of that media, and so it can have better information of its origin. This is called media-relay and it is one of the characteristics of SBC. This special aspect of SBC can be utilized as means of encryption of the media data, transcode it or intercept it. This way, the SP through the SBC can verify, can reject traffic, or limit the transmission of packets of data based on the service agreement of a given client.

Media between end-point device and SBC

Through certain protocols, like the IPSec or the SRTP, it is possible to encrypt media transmission. This is also one of the many aspects the SBC can handle. SBC can prevent certain kinds of transgressions while a transmission is happening, revising the source address and port match expected values.

4.3.6 Application Security Architecture

Part of the Trusted Domains are the AS (Application Servers). These kinds of servers are inside the local network provided by the SP but generally are located in the users geographical site. This means that they are an integral part of the Trusted Domain but they have to communicate with the external severs through an Application Security Element that belongs to the Trusted-but-Vulnerable Domain, as it is really not part of the Trusted Domain.

The Application Security Element resembles the SBC, as it also distinguishes authorized transmission from malicious ones based on what it perceives as normal versus abnormal interactions. This could mean for example excessive request volumes.

5 Availability and services

While every component of a network is necessary in order for the full system to provide a service, it is of the outmost importance for them to work together and in synchrony. All the features involved with the functions of the hardware and software may suffer from independent failures caused by internal and external sources independently. As each element is interconnected to each other in many ways, there has to be a way to measure the availability of all systems. The complete package of a network and how it performs and its reliability, taking into account all the variables included within any specific network, is what generally is referred to as a 'Carrier Grade' Network.

These failures can take the form, for example, of an algorithm not finding the operational route even if such route is available, or a certain other algorithm can not mange the traffic in a server and overloads the network. These kinds of failures point to a malfunction within the software area or the network protocols. Additionally to the software failures, there may also be problems with the hardware or error coming from an external source.

Hardware errors and failures can come from component wear-out, as it happens with most physical materials, but it can also be due to intentional attacks or even natural disasters. Materials wearing out are the most common of the hardware failures and mostly systems are designed to have this in an independent way. On the other hand, failures from the software side can take a very different pattern. Software failures come from the very beginning to the very end, meaning that every device connected to the system is running the same error, depending on whether they are using the same vendor and release.

To ensure availability of a system, Service Providers have to comply with certain minimal infrastructure designs. In this specific case, these standards are issued by the European Commission, and particularly by the FICORA within the Finnish territory. These minimal requirements needed for having optimal services is what transforms a network into a 'Carrier Class' one. This design is called security infrastructure and it is used to prevent certain kinds of failures like:

- Overloading the systems with superfluous traffic and make sure these are not interfering with processes involved with normal operation.
- Preventing against most viruses and worms that could cause system malfunctions or network element availability.
- Prevents certain actions from users that may affect system availability to be executed by both authorized and unauthorized ones.
- Make it easier to perform a quick recovery should a problem arise and there is need for one.

5.1 Availability

As stated before, we will define availability according to the European Commission guidelines contained in the *The treatment of VoIP under the EU Regulatory Framework*. [20] According to this paper, a system must comply with the following objectives set out in the EU regulatory framework are:

- Promoting competition is key, by incentivizing innovation, opening the technology markets and reduce the entry barriers for industries that have to do with VoIP.
- Expand and unify a unified market in Europe.
- To have in every moment the interest of the populations as the first priority.

With these objectives in mind, there has to be certain guidelines. In this case, we have the Public Available Telephone Service (PATS) means:

- All services have to be accessible to general public.
- Service for making and receiving national and international communications or calls.
- There must exist certain minimal services that comply with access to emergency numbers.
- There must exist national or international telephone numbering plans.

Also, there are certain steps to follow during the maintenance of a network in normal circumstances:

- Integrity and availability of the network: There must be always availability of the telephone service in case of an electric power failure.
- Emergency services: The access provided to any emergency services has to be provided by agents of PATS.
- Routing emergency calls: No information regarding the location of a call should be necessary in order to make emergency calls. Which means that location of the sender is irrelevant and the call has to be prioritised.

Some difficulties arise when certain customers change their fixed terminal to another location. If the convened place for the service is moved, the provider cannot guarantee the service, but even in those special occasions, the operator has to be capable or delivering emergency services to the customer by any means.

The services may be provided by having interconnections with the legacy PSTN, not only does this give certain flexibility to the Service Provider to deliver the necessary services, but also he could benefit from a lower termination rate.

At the same time there might also exist interconnections between VoIP networks and PSTN. VoIP users can also benefit by being able to utilize the PSTN to reach a certain user. While this might prove to be a great benefit for most users, the restriction is that the callee has to have an E.164 number.

In the specific case of Finland, a national organ enforces these rules. The bureau is called FICORA (Finnish Communications Regulation Authority). FICORA works under the general guidelines stated by the European Commission.

The rules that govern how emergency calls and lawful interception are defined in the document: *Regulation on routing and ensuring emergency traffic.* [21]

For emergency calls there are different important aspects to point out:

- Any customer must be able to access, free of charge, the universal emergency call number 112 and the police emergency number 10022.
- All emergency calls should be prioritized, meaning they have to be first in the call queue; and the routing should also have the necessary resources to redirect any emergency to the appropriate (ERC) Emergency Response Center.
- The rules regarding service operation and restrictions during an electric breakdown should be stated clearly to every user and consumer.

Communications networks are rated on the basis of their importance into different levels of priorities. These priorities mark the significance of a requirement in descending order of importance 1-5. The rating guidelines and procedures are contained in *Reasons for and application of regulation 54 on priority rating, redundancy, power supply and physical protection of communications networks and services* . [22]

This system is used as a tool to state the minimum requirements for the reliability of operation and protection of the networks used for providing communication services. The descending numerals just advises how important is a certain component within the system as a whole.

Table 1: Priority rating specification

Priority in the Regulation text	**Equivalent priority**
Very important transmission or switching system	Priority 1
Important transmission or switching system	Priority 2
Very significant network concentrator, base station, transmission system, or similar	Priority 3
Significant network concentrator, base station, transmission system, or similar	Priority 4

Priority 1: Battery backups are set as additional resources so that emergency calls will always have sufficiently power supply. This comes in addition to the other

back-up structure.

Priority 2: Uninterrupted power supply must be a critical priority. For that purpose, there always have to exist an emergency power supply to ensure the network will work during emergencies.

Priority 3: 12 hours is the minimum time for a battery backup for a communications network outside a population center. This ensures that the communications will not be interrupted even when there are longer electric power failures.

Priority 4: The only exception permitted for the 6 hours backup time for a communications network occurs when there is no possible way to meet the specified requirements at an expected reasonable cost. This will be available only if the components located outside populated areas have an estimated backup time of 12 hours.

Priority 5: The absolute minimum backup setup time allowed for a communication network is 6 hours.

Table 2: Battery back-up time

Prio.	**Battery set back-up time**	**Back-up time posted outside population center**
Prio. 1	Always	-
Prio. 2	Always	-
Prio. 3	-	12 hours
Prio. 4	6 hours	12 hours
Prio. 5	6 hours	-

Within the required parameters specified by FICORA, performance and recommended times are also included in the mix for services. In the *Explanatory notes to regulation 58 on the quality and universal service of communications networks and services* document [23], these parameters and explained and defined with in detail. For the necessary connection times on call setup FICORA requires:

- 3 seconds is the maximum allowed time for connection between two fixed networks.
- 5 seconds is the maximum allowed time for connections between a wireless network and a fixed network.
- 7 seconds is the maximum allowed time for connections between wireless networks.

Within the parameters recommended by FICORA, any service provider/operator must ensure that its physical communications networks are designed in a manner that the services can be provided without end-to-end delay of more than 150 ms

with a variation of 15 ms.

The official guidelines concerning the delay limits and its approved variation are as follows. These rules are given for the values of one-way user-to-user delay:

- less than 100 ms: Affects slightly the functioning of the conversation.
- less than 150 ms: Delay is acceptable. Good user experience.
- Between 150 and 400 ms: Delay is acceptable. Any failure in service quality must be made known to the customer.
- more than 400 ms: Delay is unacceptable. Poor user experience.

5.2 Services

VoIP can be utilized to provide many different communication services. This project is focused in offering the main VoIP communication services such as voice, video, text chat, instant messaging, presence and videoconferencing.

5.2.1 Instant Messaging and Presence

The worldwide number of users engaging in instant messaging is increasing every year as more people can access technology easier. Among the benefits offered by instant messaging, we have presence messages and notifications; but lately, the most desired features have become support for voice, video chats or file transfers. All the while, trying to make these different services work in a unified environment that could satisfy not only one but many needs. [28]

IM messages are divided into 5 categories:

- Sending and reception of text messages.
- Notifications when a user is typing or modifying a message.
- Contacts lists with presence notifications so that every user can be aware of its contact's status.
- User's pictures and the ability to upload them by the user and download them by the contacts. Also, to deliver media files directly among your list of friends.
- Control over session parameters such as logging in or out of the system.

The biggest advantage of Instant Messaging is that it gives the user a real time communication at any time or place. As opposed to email communication, instant messaging can be accessed and delivered in a timelier manner. Contact's statuses or presence is also delivered in real time. Availability is defined in other words as the actual status of a communication contact. Certain features can be attributed to presence, such as:

- The status is the base for may other kinds of communication services like chat, email or media.
- It expands the offering of services, and allows for certain new automatic ones like callbacks or conferencing.
- The new services also signify new markets for the carriers, the service providers, hosting companies, among other market players.
- These services can be integrated with other applications to expand the service potential in a unified environment.

A SIMPLE (SIP for Instant Messaging and Presence) working group of IETF has been working to develop a standardized protocol for SIP based IM communications. SIMPLE is an open standard that deals with a set of various extensions directly targeted to support the presence of IM with SIP.

The group has also been trying to ease the acceptance of this new standard protocol with the addition of certain API's (Application Programming Interface) so that as much developers as possible adopt it. Thanks to this API's IM would enrich its experience and improve its performance.

Thanks to these tools, the IM environment can be enhanced, and the more API's there are, the more developers can design and adapt different applications to different devises and operating systems. SIMPLE extends the functionalities of SIP by providing certain features that improve its IM services. The SIMPLE group also has certain recommended techniques and method to work with SIP:

- Subscribe Method: This is implemented when a certain user needs to know the presence status/information of another user.
- Notify Method: After the subscription method, the authorization notification is created. When an event occurs to which a watcher has subscribed to, the watcher will be notified using this method.
- Message Method: This method is only implemented when a certain user wants to send an IM.

Mobile communications are the best example of market growing within the area of IM as an increase in the quantity and quality of applications concerning this service has become the most sought after feature.

6 Conclusions

Voice over IP is one of many technologies that will affect the electronic communication sector over the following years. It offers the potential to increase competition, to stimulate new and innovative services for the general population, and also to the enterprise sector thanks to cost reduction and flexibility it produces to the providers and then passed to the customers. As an important benefit of the IP Telephony, the integration of voice and data applications, can result in more effective business processes and better services for everyone.

Protocols are needed for VoIP when we talk about communications based on software. Some protocols are in charge of management, others of security, distribution of resources, converting signals etc. All of them are an essential part of VoIP communications. They all have certain weaknesses and vulnerabilities, but they also will surely evolve to more modern versions and become better prepared to form more efficient infrastructures, but they are indeed critical if we are to live and communicate through IP networks.

For instance, SIP is a necessary protocol for VoIP communications but unfortunately; it suffers from certain vulnerabilities that can affect its normal functions. These vulnerabilities come in the form of how secure the call can be and how much of an inconvenience the ultimate customer is willing to sacrifice in order to have a reliable and trusted service. Every VoIP communication takes place on the Internet, which can be seen as an advantage or a threat. Making use of the many mechanisms at our disposal, the VoIP service can become a reliable option to established telephony.

In theory, the SIP security mechanism through the SIP system can guarantee the confidentiality and integrity of the VoIP service. But on the other hand, SIP security mechanisms do not take into account the availability of the system and user privacy. There are workarounds for theses inconveniences, but they only add to the many other challenges that VoIP faces.

Furthermore, technology is balancing itself to a wireless environment and with that, mobile networks are taking a special place in the minds of service providers and applications creators. On 3G networks, it is still possible to support either circuit-switched voice and packet-switched voice. That is not the case with the new and expanding 4G networks, as this will be operated on a completely based on IP. If we come to a world dominated by wireless 4G networks, legacy PSTN will become obsolete and diminish in importance when talking about voice services. This new communication environment will be in great need of IMS and services, not only access like legacy networks.

The flexibility that certain systems and its protocols are giving service providers, will expand the market also to niche applications as the related costs of entering a

certain industry (VoIP in this case) will decrease dramatically. With the addition of new Application Servers, we can augment the quantity and even the quality of the services offered trough IP networks. This service can then be distributed to any subscriber using IMS infrastructure.

The importance of SIP initiated communications may prove to be key in the near future to all VoIP transmissions. It is already been used as the standard in many systems offering communication services for private consumers, but also for big enterprises. On the other hand, H.323 based video conferencing is also becoming very common and widely deployed in the enterprise communications.

SIP is not an easy protocol to secure, its expected usage between elements with no trust at all, and its user-to-user operation makes security a difficult issue. The final security of a device or service is not achieved by securing a single protocol, it involves a complete system with which it will be implemented the solution. SIP and VoIP will probably evolve if not change completely to offer better security and reliability.

Quality of Service for voice is a critical feature for real-time Internet, therefore it is compulsory that the applications use different CAC and Precedence and Preemption mechanisms in order to better utilize the network resources and the performance of the whole system besides the emergency services needed to convey the exact location of the caller so that assistance can be dispatched to where it is required. Certain priorities will have to be established in order for the resources to be used in the most efficient but also human way possible. Every country and service provider must decide what priorities are more important.

Overall, VoIP and it multimedia counterparts will become the standard for communication in the long run.

The importance of legacy networks has not been forgotten as a good part of this thesis deals with the interoperability of the PSTN, the digital networks and the mobile ones. A good balance of new technology with the appropriate protocols in accordance to the European Directives will define the sector in the years to come. Every country will have to prove with different balances in order to benefit from the unstoppable change.

7 Future perspectives

The future of SIP will focus somehow in solving the security problems in SIP's large-scale adoption in communications services across multiple kinds of networks. Many resource intensive protocols are used to ameliorate the inconveniences of the SIP communications and VoIP.

Certainly, security and privacy will play a decisive roll when considering to add some features to the SIP protocol, but on the other hand, regulators play also a very important role in SIP's fate and development because of different laws and regulations implanted. In this case, the laws come from the European Commission, but surely these are just minimal requests compared to the overwhelming number of standards and improvements that every year come out and that cannot be regulated. The field of technology has always been faster that the creating of laws. So in this context, it is the laws that follow the advanced and quick growth of the scientific area and SIP could very well be in that category.

More specifically, the concept of SIP trunking puts companies on the path to the future of communications interworking with the support for SIP + TLS/SRTP in all of the phones and VoIP switching platforms. The list of SIP-based equipment is nowadays not large enough and it has to grow yet, but the benefits exist for an entrepreneur company to explore and take the technology further. Many years ago certain formats where not widely used, but nowadays are the standard.

One of the most important points to take into account is the money revenue for operators. Operators do not invest money in technology if that technology cannot be expected to bring sufficient revenues.

Presence and instant messaging are now mainstream products with consumers and, in the enterprise, complementing or sometimes replacing voice communications in some situations. Cost cutting and better services are pushing for a change that has already arrived to certain nations. Even for VoIP, presence has emerged not only as a valuable enhancement, but presence may be the dial tone of the twenty-first century. In the near future, the old dial-tone in telephony may well be replaced by presence information, and rich multimedia will replace the narrowband voice communications used in circuit-switched telephony.

References

[1] ITU-T. 2009. Packet-based multimedia communications systems. Recommendation H.323, International Telecommunication Union.

[2] J. Hautakorpi, G. Camarillo, R. Penfield. 2010. Requirements from Session Initiation Protocol (SIP) Session Border Control (SBC) Deployments. RFC 5853.

[3] ITU-T. 2007. Gateway control protocol: Multi-level precedence and preemption package. Recommendation H.248.44, International Telecommunication Union.

[4] H. Schulzrinne, J. Polk. 2006. Communications Resource Priority for the Session Initiation Protocol (SIP). RFC 4412.

[5] D. Goderis, H. De Neve, Y. T Joens, J. De Vriendt, T. Soetens. 2001. Towards an integrated solution for multimedia over IP. In: Alcatel Telecommunications Review 2001, pages: 97-103.

[6] J. Xue. 2007. A Framework for Military Precedence-Based Assured Services in GIG IP Networks. In: IEEE Military Communications Conference. MILCOM 2007, pages 1-7.

[7] W. Shengquan, M. Zhibin, X. Dong, Z. Wei. 2006. Design and implementation of QoS-provisioning system for voice over IP. In: Parallel and Distributed Systems, IEEE Transactions, pages 276-288.

[8] O. Salonen / Supervisor: R. Kantola. 2008. Impact of serving GPRS support node pooling on signaling. M. St. thesis, Aalto University, Finland.

[9] R. Zhao / Supervisor: S. Halme. 2002. Signalling study on interworking between UMTS IP Multimedia Subsystem and PSTN network. M. St. thesis, Aalto University, Finland.

[10] ITU, 1988. Introduction to CCITT signalling system no. 7. Recommendation Q.700, International Telecommunication Union.

[11] Z. Dawen. 2005. Distributed architecture of VOIP for firewall/NAT Traversing. In: Wireless Communications, Networking and Mobile Computing International Conference IEEE, pages 1160-1163.

[12] H. Qingyang, C. Hsiao-Hwa. 2006. A call admission control framework for voice over WLANs. In: Wireless Communications, IEEE, pages 44-50.

[13] X. Li, Z. Mi. 2002. Application of mobile agent to CAC in VoIP networks. In: TENCON '02. Proceedings. IEEE Region 10 Conference on Computers, Communications, Control and Power Engineering, pages: 736-739.

[14] S.R. Lima, P. Carvalho, V. Freitas. 2007. Admission Control in Multiservice IP Networks: Architectural Issues and Trends. In: Communications Magazine, IEEE, pages: 114-121.

[15] J. Saldana, J. Murillo, J. Fernandez-Navajas, J. Ruiz-Mas. 2011. QoS and Admission Probability Study for a SIP-Based Central Managed IP Telephony System.In: New Technologies, Mobility and Security (NTMS), 4th IFIP International Conference IEEE, pages: 1-6.

[16] Z. Zhi-Li, D. Zhenhai, T. Yiwei. 2001. On Scalable Design of Bandwidth Brokers. In: IEICE TRANS: COMMUN., pages: 2011-2020.

[17] K. Keith, M. Petros, S. Sunil, T. Rajesh, W. Larry. 2001. Bandwidth broker architecture for VoIP QoS. In: Voice Over IP (VoIP) Technology.

[18] L. Basic, U. Vizek, V. Bolt, Z. Naglic. 2004. Routing solution for VoIP calls in large-scale IP MM networks. In: Electrotechnical Conference. Proceedings of the 12th IEEE Mediterranean, pages: 673-676.

[19] L. Jin, H. Yoon, W. Sang, J. Soo. 2010. Routing mechanism for VoIP emergency calls in IP Multimedia System. In: Advanced Communication Technology (ICACT), The 12th International Conference IEEE, pages: 392-395.

[20] European Comission. 2004. The treatment of Voice over Internet Protocol (VoIP) under the EU Regulatory Framework. European Commission Information Society Directorate-General.

[21] FICORA. 2007. Application of communications legislation to VoIP services in Finland. Finnish communications regulatory authority.

[22] FICORA. 2008. On priority rating, redundancy, power supply and physical protection of communications networks and services. Finnish communications regulatory authority.

[23] FICORA. 2010. On the quality and universal service of communications networks and services. Finnish communications regulatory authority.

[24] T. Magedanz, E. Madeira. 2005. Emergency telecommunication support for IP telephony. In: IPOM'05 Proceedings of the 5th IEEE international conference on Operations and Management in IP-Based Networks.

[25] S. Mushtaq, O. Salem, C. Lohr, A. Gravey. 2008. Distributed call admission control in SIP based multimedia communication. In: International congress on Networked Electronic Media, France.

[26] K. Carlberg, R. Atkinson2004. IP Telephony Requirements for Emergency Telecommunication Service (ETS). RFC 3690.

[27] H. Sinnreich, Alan B. Johnston. 2006. Internet Communications using SIP, Wiley Publishing.

[28] V. Hilt, I. Widjaja. 2008. Controlling Overload in Networks of SIP Servers. Bell Labs, Alcatel-Lucent, IEEE, pages: 83-93.

[29] W. Changzhou, W. Guijun, W. Haiqin. 2006. Quality of Service (QoS) Contract Specification, Establishment, and Monitoring for Service Level Management. In: 10th IEEE International Enterprise Distributed Object Computing Conference Workshops (EDOCW'06).

[30] G. Camarillo. 2011. A Service-enabling Framework for the Session Initiation Protocol (SIP). Ph. D. thesis. Aalto University, Finland.

[31] H. Khlifi, J. Gregoire. 2007. A Modular Architecture for Providing Carrier-Grade SIP Telephony Services. In: Third IEEE International Conference on Wireless and Mobile Computing, Networking and Communications.

[32] M. Balasaygun, R. Steiner. 2010. Signaling using Binary form of SIP Messages. In: Patent Application Publication, US.

[33] F. Ricciato. 2004. Measurement-based Optimization of the GPRS/UMTS Core Network. In: Technical Report FTW-TR-2005-009.

[34] L. Mokdad, M. Sene, A. Boukerche. 2011. Call Admission Control Performance Analysis in Mobile Networks Using Stochastic Well-Formed Petri Nets. In: IEEE Transactions on Parallel and Distributed Systems, pages: 1332-1341.

[35] B. Todtmann, E. Rathgeb. 2007. Advanced Packet Filter Placement Strategies for Carrier-Grade IP-Networks. In: 21st International Conference on Advanced Information Networking and Applications Workshops (AINAW'07), pages:415-423.

[36] W. Greene, B. Lancaster. 2006. Carrier-Grade: Five Nines, the Myth and the Reality. In: Pipeline Volume 3, Issue 11.

[37] A. Pang, Y. Lin, W. Chen. 2005. An IPv4-IPv6 Translation Mechanism for SIP Overlay Network in UMTS All-IP Environment. In: IEEE Journal on selected areas in communications, pages: 2152-2160.

[38] M. Martin, P. Hung. 2005. Towards a security policy for VoIP applications. In: Canadian Conference on Electrical and Computer Engineering, IEEE, pages: 65-68.